DOCUMENT RESUME

ED 445 914 SE 064 086

AUTHOR	Klein, Stephen; Hamilton, Laura; McCaffrey, Daniel; Stecher, Brian; Robyn, Abby; Burroughs, Delia
TITLE	Teaching Practices and Student Achievement: Report of First-Year Findings from the 'Mosaic' Study of Systemic Initiatives in Mathematics and Science.
INSTITUTION	Rand Corp., Santa Monica, CA.
SPONS AGENCY	National Science Foundation, Arlington, VA.
ISBN	ISBN-0-8330-2879-0
PUB DATE	2000-00-00
NOTE	104p.
CONTRACT	ESI-96-15809
AVAILABLE FROM	RAND, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138; web site: http://www.rand.org
PUB TYPE	Books (010) -- Reports - Descriptive (141)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Academic Achievement; Critical Thinking; *Educational Change; Elementary Secondary Education; Mathematics Education; Problem Solving; Professional Development; Science Education; *Teaching Methods; Urban Schools
IDENTIFIERS	National Science Foundation; *Systemic Educational Reform

ABSTRACT

This book reports on the first wave of the Mosaic Project which investigates the effects of educational change and teaching practices in mathematics and science on student achievement. The Mosaic project was supported by the National Science Foundation (NSF) and featured two waves. The states that participated in the first wave were State Systemic Initiatives (SSIs) in Connecticut and Louisiana; Urban Systemic Initiatives (USIs) in Columbus, OH and San Francisco, CA; and Local Systemic Change (LSC) projects in Fresno, CA, and El Paso, Socorro, and Ysleta, TX. This document contains four chapters: (1) "Introduction: The Systemic Initiatives Programs, Earlier Evaluations of Systemic Initiatives, Evidence of Relationships between Teaching Practices and Achievement, Measuring Student Achievement, Overview of the Mosaic Project"; (2) Methods: Site Selection, School, Subject, and Grade-Level Selection, Data Collection, Participation Rates, Analysis"; (3) First-Year Results: Distribution of Teaching Practices, Relationships between Teaching Practices and Achievement, Alternative Formulations for Site Models, Differences between Test Formats, Study Limitations"; and (4) Discussion: Summary of Year 1 Findings, Plans for Future Data Collection and Analysis". (Contains 33 references.) (YDS)

Reproductions supplied by EDRS are the best that can be made

ED 445 914

Teaching Practices and Student Achievement: Report of First-Year Findings from the "Mosaic" Study of Systemic Initiatives in Mathematics and Science.

SE 064 086

Stephen Klein
Laura Hamilton
Daniel McCaffrey
Brian Stecher
Abby Robyn
Delia Burroughs

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

RAND

EDUCATION

Teaching Practices and Student Achievement

Report of First-Year Findings from the 'Mosaic' Study of Systemic Initiatives in Mathematics and Science

Stephen Klein, Laura Hamilton,
Daniel McCaffrey, Brian Stecher,
Abby Robyn, Delia Burroughs

3

The research described in this report was supported by the National Science Foundation, grant number ESI-96-15809.

ISBN: 0-8330-2879-0

Building on more than 25 years of research and evaluation work, RAND Education has as its mission the improvement of educational policy and practice in formal and informal settings from early childhood on.

RAND is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND® is a registered trademark. RAND's publications do not necessarily reflect the opinions or policies of its research sponsors.

© Copyright 2000 RAND

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from RAND.

Published 2000 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1333 H St., N.W., Washington, D.C. 20005-4707
RAND URL: http://www.rand.org/
To order RAND documents or to obtain additional information, contact Distribution Services: Telephone: (310) 451-7002; Fax: (310) 451-6915; Internet: order@rand.org

Teaching Practices and Student Achievement

Report of First-Year Findings from the 'Mosaic' Study of Systemic Initiatives in Mathematics and Science

Stephen Klein, Laura Hamilton,
Daniel McCaffrey, Brian Stecher,
Abby Robyn, Delia Burroughs

RAND
EDUCATION

PREFACE

During the 1990s, the National Science Foundation (NSF) funded a number of large-scale initiatives designed to change the way mathematics and science are taught in schools. These efforts, called Systemic Initiatives (SIs), shared a common emphasis on aligning all aspects of the educational system in support of ambitious curriculum and performance standards. Particular emphasis was placed on teacher training and professional development to promote changes in instructional practice that would enable students to achieve the new standards.

Funds were given to states, to urban school districts, and to consortia of districts to implement reforms consistent with NSF's purposes. Sites had considerable flexibility in designing their programs, and they adopted very different strategies for promoting reform. As a result, initial research on the SIs focused on the complex process of development and implementation. Although individual sites gathered information, after five years of funding, NSF had no broad picture of the effects of the reform on student achievement.

In 1996, NSF provided funds to RAND to investigate the relationships between student achievement in mathematics and science and the use of instructional practices that are consistent with systemic reforms. The study, called the Mosaic project, was conducted in two waves: A set of six sites (including both states and urban districts) that were implementing systemic reforms was studied during the 1996–97 school year, and a similar set of six sites was studied during the 1997–98 school year. The same basic analytic design was replicated at each site, and the study draws much of its power and generalizability from this replication.

6

This report presents results for the first wave of the study. The results should be of interest to educational policymakers at all levels of government, as well as to program developers and school administrators interested in mathematics and science education.

CONTENTS

Preface iii

Figures vii

Tables ix

Summary xi

Acknowledgments xvii

Chapter One

INTRODUCTION 1

The Systemic Initiatives Programs 2

Earlier Evaluations of Systemic Initiatives 3

Evidence of Relationships Between Teaching Practices and Achievement 5

Measuring Student Achievement 5

Overview of the Mosaic Project 6

Chapter Two

METHODS 9

Site Selection 9

School, Subject, and Grade-Level Selection 10

Data Collection 10

Student Achievement Data 11

Teacher Questionnaires 13

Demographic Data 15

Participation Rates 16

Analysis 16

Chapter Three
FIRST-YEAR RESULTS 19
Distributions of Teaching Practices 19
Relationships Between Teaching Practices and Student Achievement 21
Alternative Formulations for Site Models 30
Differences Between Test Formats 30
Study Limitations 34

Chapter Four
DISCUSSION 37
Summary of Year 1 Findings 37
Plans for Future Data Collection and Analysis 39

Appendix
A. Participation at Year 1 Sites 41
B. Items on Teaching-Practices Scales 47
C. Full Regression Models 51
D. Details of Pooled Analysis of Regression Coefficients 65
E. Results from Analysis of Format Differences 67
F. Sensitivity Analyses: Use of Contemporaneous Test Scores 69
G. Sensitivity Analyses: Combining Reform and Traditional Scales in a Single Model 73

References 81

FIGURES

S.1. Pooled Results from Mathematics Analyses xv
S.2. Pooled Results from Science Analyses xv
3.1. Distribution of Teacher Scores on Reform and Traditional Scales by Site and Subject 21
3.2. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Open-Ended Tests 23
3.3. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Multiple-Choice Tests 23
3.4. Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Open-Ended Tests . . . 24
3.5. Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Multiple-Choice Tests 24
3.6. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Open-Ended Tests 25
3.7. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Multiple-Choice Tests 25
3.8. Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Open-Ended Tests 26
3.9. Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Multiple-Choice Tests 26

3.10. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Difference Between Open-Ended and Multiple-Choice Tests 32
3.11. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Difference Between Open-Ended and Multiple-Choice Tests 32
3.12. Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Difference Between Open-Ended and Multiple-Choice Tests 33
3.13. Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Difference Between Open-Ended and Multiple-Choice Tests 33
F.1. Estimated Coefficients for Mathematics and Science Tests, Sites 2, 3, 4, and 6: Models Include Prior-Year, Contemporaneous, or No Test Scores 71
G.1. Estimated Coefficients for Mathematics and Science Tests 74
G.2. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Open-Ended Tests 75
G.3. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Multiple-Choice Tests 75
G.4. Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Open-Ended Tests 76
G.5. Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Multiple-Choice Tests . . 76
G.6. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Open-Ended Tests 77
G.7. Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Multiple-Choice Tests 77
G.8. Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Open-Ended Tests 78
G.9. Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Multiple-Choice Tests 78

TABLES

S.1. Details on Participating Sites xiii
S.2. Standardized Regression Coefficients from Pooled Analyses of the Relationship Between Instructional Practices and Student Achievement xiv
2.1. Details of Year 1 Participating Sites 11
2.2. Sites, Subjects, and Assessment Instruments 12
3.1. Coefficients from Pooled Analyses of Relationships Between Instructional Practices and Achievement . . . 28
B.1. Items on Reform-Practices Scale for Mathematics . . . 47
B.2. Items on Traditional-Practices Scale for Mathematics 48
B.3. Items on Reform-Practices Scale for Science 49
B.4. Items on Traditional-Practices Scale for Science 50
C.1. Site 1 Regression Models for Mathematics Tests (Grade 4) 51
C.2. Site 2 Regression Models for Science and Mathematics Tests (Grade 5) 53
C.3. Site 3 Regression Models for Science and Mathematics Tests (Grade 5) 56
C.4. Site 4 Regression Models for Science Tests (Grade 5) 60
C.5. Site 5 Regression Models for Science and Mathematics Tests (Grade 7) 62
C.6. Site 6 Regression Models for Science and Mathematics Tests (Grade 7) 64
D.1. Results from Pooled Analyses of Relationships Between Practices and Achievement 66
E.1. Standardized Coefficients for Models Predicting Differences Between Formats 67

E.2. Results from Pooled Analyses of Differences Between Formats 68
G.1. Results from Pooled Analyses of Relationships Between Practices and Achievement 79

SUMMARY

During the 1990s, the National Science Foundation (NSF) supported the efforts of several states and large school districts to change the way mathematics and science were taught. These programs, called Systemic Initiatives (SIs), emphasized aligning all aspects of the educational system with ambitious curriculum and performance standards. The funded sites had considerable flexibility in designing their programs, and they used many different strategies to promote reform. However, extensive in-service training for teachers was often the centerpiece of their efforts.

In 1996, NSF awarded RAND a grant to investigate a key assumption underlying the SI program, namely, that greater use of instructional practices that are aligned with the reform would lead to improved student achievement in mathematics and science. To carry out this research, RAND and NSF collaborated in identifying 11 SI sites across the country that were emphasizing reforms in mathematics, science, or both. Data were collected at the following six sites in the first year: Fresno, CA; San Francisco, CA; Connecticut; Louisiana; Columbus, OH; and the combination of El Paso, Socorro, and Ysleta, TX.

RESEARCH DESIGN

The same basic research design was used at all sites. This design had the following three major components: (1) a measure of instructional practices, (2) assessment of student achievement, and (3) an analysis of the strength of the relationship between instructional practices and student achievement after controlling for student background characteristics.

The teacher questionnaire used to measure instructional practice was administered to a large sample of the teachers at each site who taught mathematics and/or science at the grade level being studied at that site. The questionnaire asked teachers about the frequency with which they used various types of reform and traditional instructional practices. For example, it asked how often the students conducted their own science experiments (reform) versus listening to the teacher lecture (traditional). Horizon Research, Inc. (HRI), under a subcontract from RAND, had primary responsibility for designing and validating this questionnaire.

The student assessment component involved administering tests in mathematics and/or science to a large sample of students at each site's targeted grade level. To conserve resources and reduce the testing burden on students and teachers, scores from existing statewide or districtwide assessment programs were used when such scores were available. We augmented these "local" measures with tests administered by RAND staff and consultants. Students at all but one site took both a multiple-choice and an open-response test in the subject(s) assessed at their school. Consequently, the specific tests used at each site differed.

The third research design component common to all sites involved the statistical methods used to analyze the data. The relationship between instructional practices and student achievement was examined, after controlling for relevant student background characteristics (such as prior-year test scores and whether the student was in a free or reduced-price lunch program). However, the specific control variables differed somewhat across sites (e.g., not all sites had test scores from a prior year). The analyses also controlled for the "nesting" of students within classrooms. Taken together, these controls helped to isolate and measure the relationships between use of the instructional practices and student achievement.

Table S.1 shows the grade level and subject(s) studied at each first-year site and the number of schools, teachers, and students that participated in the study.

TEACHER QUESTIONNAIRE DATA

Analyses of the teacher questionnaire revealed that the frequency with which a teacher used the reform practices was generally un-

Table S.1

Details on Participating Sites

Site	Grade	Subject	Number of Schools[a]	Number of Teachers	Number of Students (varied by test in some sites)
1	3rd	Math	17	46	804
2	5th	Math	20	100	1,651–1,686
2	5th	Science	20	99	1,639–1,662
3	5th	Math	18	73	1,366–1,451
3	5th	Science	20	74	1,367–1,438
4	5th	Science	19	45	909–932
5	7th	Math	17	48	2,937–3,018
5	7th	Science	19	33	2,047–2,079
6	7th	Math	25	57	3,237
6	7th	Science	25	52	3,279

[a]Some schools at each site are included in both the mathematics and science samples.

related to the frequency with which that teacher used traditional practices. For example, some of the teachers who used the reform practices relatively frequently also used traditional practices frequently, while other frequent users of the reform practices used the traditional practices only rarely. Thus, the two aspects of practice were not opposite ends of a single dimension.

The analyses also showed substantial variability in instructional practices within schools, regardless of the degree of implementation of the reform program. There are many plausible explanations for this finding—for example, not all teachers were trained in the same way or at the same time—but examining the source of this relatively large within-school variation in instructional practices was beyond the scope of our study.

STUDENT ACHIEVEMENT

After controlling for student background characteristics, we found a generally weak but positive relationship between the frequency with

which a teacher used the reform practices and student achievement. This relationship was somewhat stronger when achievement was measured with open-response tests than with multiple-choice tests. The use of traditional practices was generally unrelated to achievement. The foregoing trends held for both mathematics and science; and they were generally consistent across the six sites, i.e., in most cases, the pooled results across sites were not driven by the data at one or two sites.

Table S.2 illustrates these trends by contrasting the standardized regression coefficients for the relationship between student achievement and the teacher-reported frequency of using reform and traditional instructional practices. These trends are also illustrated by Figures S.1 and S.2, which show the pooled (across-site) effect sizes (i.e., the increase in student achievement, as measured in standard-deviation units, that corresponds to a one-standard-deviation increase on the instructional-practices scale).

DISCUSSION AND CONCLUSIONS

Taken together, the data in Table S.2, the results in Figures S.1 and S.2, and the consistency of findings across sites provide some (albeit weak) support for the hypothesis that the reform practices are associated with improved student achievement in both mathematics and science. However, as with most large-scale field studies, there are many factors that may have artificially increased or decreased the observed effect sizes.

Table S.2

Standardized Regression Coefficients from Pooled Analyses of the Relationship Between Instructional Practices and Student Achievement

	Mathematics		Science	
Instructional Practices	Multiple-Choice	Open-Response	Multiple-Choice	Open-Response
Reform	0.030	0.053[a]	0.045[a]	0.079[a]
Traditional	0.001	–0.025	0.006	–0.018

[a]Statistically significant relationship ($p < 0.05$).

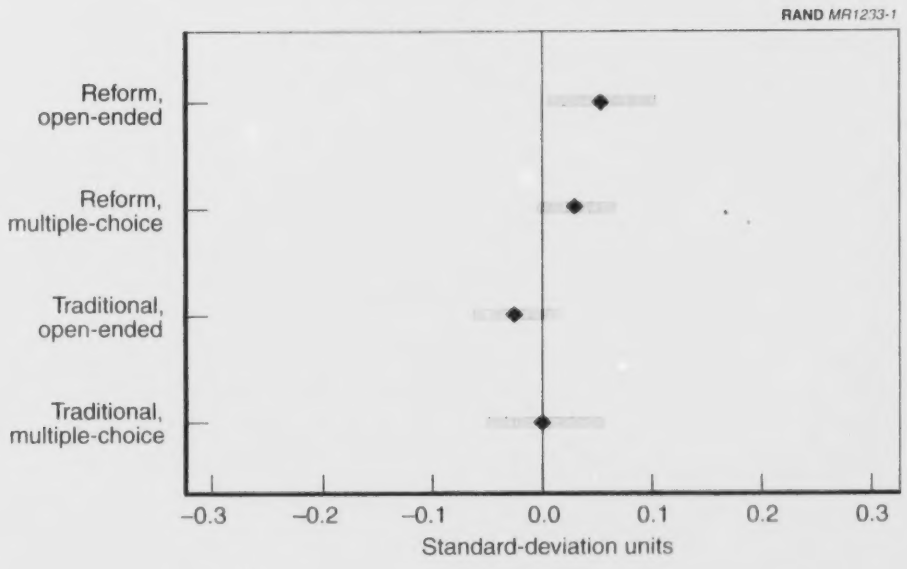


Figure S.1—Pooled Results from Mathematics Analyses

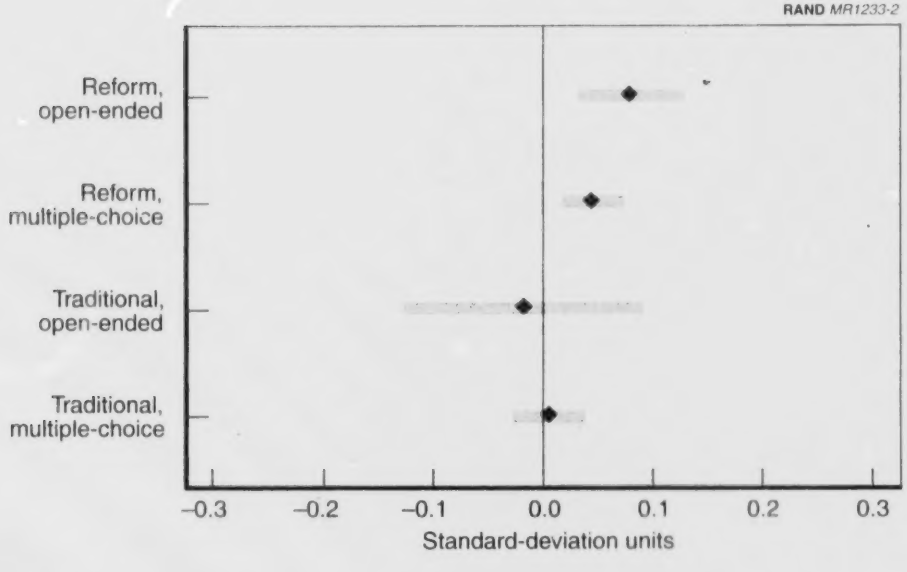


Figure S.2—Pooled Results from Science Analyses

The following examples illustrate the problem: Teachers may not always have provided accurate reports of the extent to which they used various instructional practices, and some may not have become proficient in the use of the reform practices at the time the data were collected. The tests used to measure student achievement may not have been aligned especially well with the reform curriculum. Students whose teachers use the reform practices relatively frequently may differ from other students for reasons that are unrelated to the use of the reform practices per se. Finally, students may have to experience the reform practices for more than one year in order for these practices to have a significant impact on achievement. Nevertheless, the consistency of results across sites, despite the differences among sites (e.g., in the grade levels, control variables, and tests used), is encouraging. Data from the second year of the study will provide additional evidence to aid in the interpretation of these findings.

ACKNOWLEDGMENTS

The Mosaic project was supported by the National Science Foundation, Division of Elementary, Secondary, and Informal Education, under grant ESI-96-15809. The project would not have been successful without the cooperation and support of key individuals at each of the participating sites. The first-year sites included State Systemic Initiatives (SSIs) in Connecticut and Louisiana; Urban Systemic Initiatives (USIs) in Columbus, Ohio, and San Francisco, California; and Local Systemic Change (LSC) projects in Fresno, California, and El Paso, Socorro, and Ysleta, Texas. We are indebted to the individual leaders of the reform program at each of these sites, as well as to the regular state and district administrations. The list of those who contributed directly to this project includes the following individuals:

Fresno, California

Charles E. McCully, Superintendent
Georgina Takemoto, PI, Assistant Superintendent
Robert Grobe, Administrative Analyst
Paul Messenhiemer, Data Systems
Linda Ball, USI, LSC
Bob Harringon, Assistant Superintendent for Research and Evaluation

San Francisco, California

Waldemar Rojas, Superintendent
Maria Santos, Associate Superintendent
Pat Anderson, Supervisor, Testing and Evaluation

20

Sandra Lam, Program Director
Carmelo Sgarlato

Connecticut

Richard Cole, Executive Director, Connecticut Academy for Education
Douglas Rindone, Chief, Bureau of Research and Student Assessment, State Department of Education
Steve Leinwand, State Mathematics Consultant
Mari Muri, State Mathematics Consultant
Steve Weinberg, State Science Consultant
Bob Rosenbaum, PIMMS Department

Louisiana

Cecil Picard, State Superintendent
Kerry Davidson, Project Director, LaSIP
John Wallin, LaSIP
Carl Frantz, Coordinator for Evaluation, LaSIP and LaCept
Richard Anderson, LaSIP
Faimon Roberts, Assistant Director for Science, LaSIP
Barbara Andrepointe, Management Information Systems, Louisiana Department of Education

Columbus, Ohio

Rosa Smith, Superintendent, Columbus Public Schools
Camille Nasbe, Director, Columbus USI
Saundra Brennan, Project Evaluator, Columbus USI
Burt Wiser, Director of Assessment and Testing, Columbus Public Schools
Larry Sullivan, National Association of School Psychologists
Marc Foor, Columbus Public Schools
Maxine Smith, Site Coordinator

El Paso, Texas

Kenneth George, Interim Superintendent, El Paso Independent School District
Don Schulte, Superintendent, Socorro Independent School District

Irma Trujillo, Interim Superintendent, Ysleta Independent School District
Susana Navarro, Executive Director, El Paso Collaborative for Academic Excellence
Gabriel Della-Piana, Director of Evaluation, El Paso Collaborative for Academic Excellence
H. Susan Schneider, Evaluator
Anna Bone, Researcher
Ray Reynosa, District USI Director for El Paso ISD
Gary Ivory, Director of Research, Testing, and Evaluation
Maria Gutierrez, District USI Director for for Ysleta ISD
Joe Bob Shook, Director of Secondary Education
Joyce Zarowsk, District USI Director for Socorro ISD
Cathe Lester, Site Coordinator

We also wish to thank the hundreds of teachers, principals, and support staff who participated in the study and the thousands of students who completed our assessments, without whom this complex project could never have been possible.

Horizon Research, Inc., was responsible for the design, production, and scoring of the teacher surveys. We are grateful to Iris Weiss, Jon Supovitz, and the staff of Horizon for their careful and efficient work. CRB Associates provided data collection support in Connecticut, and Hugh and Connie Bruckerhoff deserve credit for their efforts on behalf of the Mosaic project.

Special thanks go to Janice Earle, NSF, for her encouragement, constant support, and timely assistance. There would be no Mosaic project without her foresight and guidance.

A project of this size and complexity requires substantial operational support, and we were fortunate to have the analytical skills of Tor Ormseth and the logistical assistance of Robert Reichardt and Peter Scott, all of RAND. Sharon Koga, Helen Rhodes, and Donna White contributed to the preparation of this report. Finally, the thoughtful, incisive reviews provided by Becky Kilburn and Larry Hanser improved the quality of this report.

Chapter One

INTRODUCTION

Many of the mathematics and science education reforms that are currently under way in the United States seek to improve achievement by fostering classroom practices designed to enhance the development of critical thinking and problem-solving skills, particularly among low-income and minority students. One approach being widely implemented today is called *systemic reform* because it attempts to align all parts of the educational system—curriculum, instruction, assessment, teacher preparation, and state and local policies such as graduation requirements—to promote change in the classroom and, ultimately, improve student performance (Smith and O'Day, 1991). Systemic reform efforts resulted in part from the observation that addressing one component of the educational system tended to be ineffective due to constraints imposed by other parts of the system (Hill, 1994; Knapp, 1997).

This report presents results from the first year of a study designed to investigate relationships between student achievement in mathematics and science and the use of instructional practices that are consistent with systemic reforms. We begin with background information on the reforms, particularly the initiatives currently being funded by the National Science Foundation (NSF). We summarize existing evidence on the effectiveness of these efforts and the difficulties researchers face in measuring relevant student outcomes. We then describe our approach to studying the problem, including our samples, measures, and methods of analysis. Following that, we present the results from the six sites at which we collected data during this phase of the project. The conclusion of the report summarizes

our major findings, discusses the limitations of the analysis, and suggests directions for future research and evaluation.

THE SYSTEMIC INITIATIVES PROGRAMS

In 1990, NSF launched a series of initiatives designed to promote standards-based systemic reform of mathematics and science education. Through its Statewide Systemic Initiatives (SSI) program, NSF awarded grants to 25 states and the Commonwealth of Puerto Rico from 1991 to 1993. The state level was chosen as the initial target, in part because NSF viewed state policymakers as being uniquely able to influence all aspects of the educational system, including teacher training in institutions of higher education. Grants were awarded for a five-year period, but some states were able to renew their grants for additional years.

The Urban Systemic Initiatives (USI) program, established in 1993, targets cities where large numbers of children live in poverty. This program has funded 20 large urban districts with awards of up to $15 million over five years. The program is described as a "comprehensive and systemic effort to stimulate fundamental, sweeping, and sustained improvement in the quality and level of K–12 science, mathematics, and technology (SMT) education" (Williams, 1998, p. 7). The Local Systemic Change (LSC) program was created in 1995 to fund district-based teacher enhancement through curriculum implementation at more than 50 sites. These projects are also of five years' duration, but they are typically smaller in scope, with funding based on the number of teachers served. Most of the projects are receiving between $2 million and $6 million over the five-year funding period. A Rural Systemic Initiatives (RSI) program operating in several sites completes the set of NSF initiatives. Together, these Systemic Initiative (SI) programs have received approximately $100 million per year in NSF funding. In addition, most sites supplemented their NSF grants with additional local contributions—sites are currently using Title I funds, corporate donations, and grants from private foundations to support and expand their SIs (Williams, 1998).

Although these programs vary in scope and emphasis, all are relatively long-term (five years, with a small number of SSIs being extended for an additional five years), and all attempt coordinated

reform, aligning various parts of the educational system with one another. These initiatives, in theory at least, generally involve the development of ambitious curriculum and performance standards and the mobilization of all components of the system to support and enable all students to reach those standards (Consortium for Policy Research in Education, 1995a).

To be effective, these reforms must ultimately be adopted by teachers and must take hold in the classroom (Tyack and Cuban, 1995). Thus, a primary emphasis of the SIs involves promotion of teaching practices that are assumed to facilitate student learning. Most initiatives offer professional development to teachers, and this component constitutes a fairly large proportion of the budget. For example, the SSI sites spent nearly one-third of their first-year budgets on in-service training for teachers, more than on any other category of spending (Shields, Corcoran, and Zucker, 1994). Most of this training is intended to increase teachers' use of classroom practices that are believed to improve achievement.

The kinds of practices being promoted by NSF, as well as by numerous other agencies and reformers, are consistent with curriculum standards and guidelines that have been published by the National Research Council (1996), the American Association for the Advancement of Science (1993), and the National Council of Teachers of Mathematics (1989). Common to all of these documents is an emphasis on instruction that engages students as active participants in their own learning and that enhances the development of complex cognitive skills and processes. Specific practices that are endorsed include cooperative learning groups, inquiry-based activities, use of materials and manipulatives, and open-ended assessment techniques. All of these practices are intended to support active rather than passive learning, to promote the application of critical thinking skills, and to provide opportunities to apply mathematics and science learning to real-world contexts.

EARLIER EVALUATIONS OF SYSTEMIC INITIATIVES

Numerous evaluations have been conducted by the individual SI sites and by outside organizations. Most of these evaluations have

focused on the degree of implementation of the reforms (e.g., type and frequency of professional development offered to teachers, level of participation among teachers) rather than on student outcomes. However, NSF and the sites are becoming increasingly concerned about student achievement. Many of the SI sites have reported improvement in student test scores (Williams, 1998), but most offer little if any evidence that ties this improvement directly to SI participation.

A large-scale study conducted by SRI International revealed small but statistically significant differences in test scores that favored participating over nonparticipating schools at four of seven SSI sites (Laguarda, 1998). However, this study had a number of limitations. First, the analyses did not control for any preexisting differences in the teachers and students in SSI and non-SSI schools. We have observed that sites often implement large-scale reforms in phases, and those schools that participate in the earlier phases differ in important ways (e.g., in the experience of teachers and in the socioeconomic backgrounds of students) from those that participate in later phases. Second, the analyses did not examine the degree of implementation of the reforms within schools. The fact that a school is considered part of the reform effort does not guarantee that all the teachers in the school are responding in the intended manner. Other researchers have found that teachers' use of reform practices is influenced by many factors, including the nature and frequency of professional development participation (Cohen and Hill, 1998; Weiss et al., 1998) and the degree to which the teachers understand the subject matter (Cohen and Ball, 1990). Third, the data were collected and analyzed by site personnel rather than by the external evaluators, and no effort was made to address differences in the quality of these procedures across sites.

The absence of good evaluation data on SI programs has led some policymakers to express skepticism about the value of these programs (Fox, 1998). Others have called for more-rigorous evaluations that focus on student achievement and relate it to the degree of implementation of the reforms. There is some evidence of a positive relationship between the practices promoted by the SIs and student achievement in mathematics and science, and we review this evidence below.

EVIDENCE OF RELATIONSHIPS BETWEEN TEACHING PRACTICES AND ACHIEVEMENT

If the SIs do improve student achievement, it is undoubtedly due in large part to what occurs in the classroom. For this reason, professional development and the promotion of good instructional practices are critical to the success of the initiatives. Research provides some evidence of the effectiveness of some of the individual practices endorsed by the reforms. An experiment conducted by Ginsburg-Block and Fantuzzo (1998), for example, showed that low-achieving elementary students who were placed in problem-solving or peer-collaboration situations achieved higher mathematics scores and reported higher levels of motivation than did students who received neither of these interventions. Several other studies have also demonstrated the value of peer tutoring and collaboration (e.g., Fantuzzo, King, and Heller, 1992; Greenwood, Carta, and Hall, 1988; Webb and Palincsar, 1996), as well as the benefits of contextualizing instruction in real-world problems (Verschaffel and De Corte, 1997).

A few studies have focused on relationships between student achievement and teachers' use of combinations of these practices. Cohen and Hill (1998) studied teacher-reported use of several practices consistent with the 1992 California Mathematics Framework and found that frequency of use was positively related to scores on the California Learning Assessment System (CLAS) mathematics test at the school level, after controlling for demographic characteristics. The set of teaching practices examined in that study was similar to the sets being advocated and supported by the SIs. Mayer (1998) found small positive or null relationships between a similar set of practices and student scores on a standardized multiple-choice test. Thus, there is some evidence that, in certain contexts at least, use of reform practices is related to higher student achievement.

MEASURING STUDENT ACHIEVEMENT

One difficulty in evaluating ongoing programs in general, and the SIs in particular, is a lack of appropriate measures of student achievement. Most states do not currently administer tests that are well aligned with the systemic reforms (Consortium for Policy Research in Education, 1995b). Part of this misalignment may arise because

many large-scale tests (whether developed by the state or by commercial publishers) rely on multiple-choice items, a format that does not always lend itself to measuring many of the scientific inquiry and mathematical problem-solving skills encouraged by the SIs. In addition, many state testing programs predate both the systemic reforms and the current national standards. As of 1995, 21 states did not test students in science at all (Bond, Braskamp, and van der Ploeg, 1996), although the number of states that test in science has increased in recent years.

An additional problem with state testing programs is that most do not provide data that can be used to track progress over time. In many states, students are tested only at selected grade levels (e.g., fourth, eighth, and tenth). Changes in scores of successive cohorts of students confound the effects of reforms with differences among the groups of students. In addition, improvements in scores over time, which are often cited as evidence of beneficial effects of reforms on student learning, may in many cases reflect inappropriate narrowing of the curriculum or teaching to the test (Koretz and Barron, 1998; Koretz et al., 1991). This problem is especially likely to occur when the tests are part of a high-stakes accountability system or when the same form of a test is administered multiple times. For all of these reasons, it is desirable to supplement existing state tests with additional measures whenever possible.

OVERVIEW OF THE MOSAIC PROJECT

The Mosaic project, described in this report, was designed to examine the relationship between teaching practices and student achievement in mathematics and science, a relationship that is at the heart of the SIs. We gathered data from a variety of SI sites using multiple measures of achievement to produce a "mosaic" of evidence about this relationship. Our approach is to model this relationship rather than to compare directly the performance of students whose teachers participated in different phases of the reform. One of the advantages of this approach is that we can include measures of student demographic characteristics as well as prior achievement in the model. This allows us to adjust our analyses for some of the major differences among students assigned to different classrooms and schools. Another advantage is that we measure directly the degree to

which teachers actually use both traditional and reform practices, so we can focus on instruction at the classroom level rather than at the school level. Although at many sites the school is the unit of participation in the reform, we have found that there is substantial variation in teaching practices among teachers within a school, even though teachers may have been exposed to the same training. By collecting data on individual teachers, we can address these differences in our analysis. Finally, we measure student achievement using both multiple-choice and open-response tests, including some hands-on science tasks that we developed and administered ourselves. This provides greater sensitivity to potential gains in skills than would be provided by multiple-choice tests alone, and it gives us an opportunity to explore differences between the multiple-choice and open-response formats. The Mosaic study is being conducted at 11 sites, which will provide a strong test of the strength of the relationship across sites.

We adopted this modeling approach for a number of reasons. First, it was difficult to judge the effectiveness of the comprehensive SIs directly, because the reforms were already well under way when this study began. It was impossible to collect baseline data and other information that would be necessary to evaluate the cumulative impact of the reforms on student achievement. Second, the reforms were not implemented with an outcome analysis in mind, and in general, the sites did not address research design issues when they developed their programs. For example, some sites provided training to all teachers the first year, leaving no untreated classes to use for comparative analyses. Other sites that implemented reforms in phases defined those phases on the basis of geographic region. As a result, student demographic factors were not the same for each wave of the reform, and direct comparisons between phases would not be appropriate. In addition, few sites collected any measures of teaching, so it was impossible to know whether the training actually led to differences in classroom practices.

It is important to understand that the Mosaic project is not a comprehensive evaluation of the systemic reform initiatives. These initiatives are multifaceted, multiyear efforts to bring about changes in classroom practice and in other aspects of the educational system. The reform sites have adopted a wide range of strategies to recruit and train teachers in new methods, to implement new curricula, to

provide appropriate materials, to encourage and sustain change at the school level, and to instill greater interest in mathematics and science. Their success at these tasks is the subject of a comprehensive evaluation being undertaken by SRI International (Corcoran, Shields, and Zucker, 1998; Shields, Marsh, and Adelman, 1998).

Chapter Two

METHODS

We collected data from six of 11 sites in year 1 (the focus of this report), and we recently completed data collection at six sites in year 2 (one of the sites is providing data in both years). Most of the data collection took place in elementary schools and middle schools because the bulk of the reform activities occurred at these grade levels. Our specific procedures for site selection, subject and grade-level selection, and data collection are described in this chapter.

SITE SELECTION

We knew that it would be difficult to study the relationship between reform instructional practices and achievement in the absence of a reasonable degree of reform, so we selected sites in a way that maximized the probability of encountering substantial numbers of teachers using reform practices. NSF proposed sites at which reforms in science and/or mathematics instruction appeared to be occurring, based on information drawn from their site visits and from progress reports submitted by the grantees.

Mosaic project staff visited each proposed site to discuss the goals of the study, data collection requirements, the availability of data, student achievement measures, requirements for linking teacher and student data, and local site coordination. On the basis of these visits, all of the proposed sites except three were included in the study. One proposed site declined to participate because its program was not yet advanced enough to study; the reforms at another site were so widespread in the district that the necessary variation in teaching practice was unlikely to be found, although this site was later incor-

porated into a statewide site; and the third site was excluded because the sample size was too small and testing was limited. The other proposed sites agreed to participate and to provide the necessary student, teacher, and demographic data. A local coordinator responsible for testing arrangements was designated by each site's administrators.

The six sites in our sample consisted of two states and four urban systems (three of which were single districts, and one of which was a group of three districts located in the same city). Data were collected from 324 mathematics teachers at 97 schools, and from 303 science teachers at 103 schools. At five of the sites, students received both multiple-choice and open-response and/or hands-on tests in the targeted subject(s). At the other site, multiple-choice assessments were administered, but we were unable to schedule any open-response assessments because of time constraints.

SCHOOL, SUBJECT, AND GRADE-LEVEL SELECTION

School district and program staff at each site specified the grade level(s) and subject(s) in which they believed reform practices were most pervasive, then nominated schools to participate in the study. The same basic research design was used at each site. We asked local staff to select approximately 10 schools in which there was good reason to believe mathematics and/or science instruction reforms had been implemented, and 10 demographically similar schools in which reforms had yet to be implemented. (All of the sites had been involved in the reform for more than one year, but some had not yet implemented the reforms in all of their schools.) We used the nominations only to ensure variation in teaching practices; we did not compare the high- and low-implementing schools with one another directly. Table 2.1 lists the grade(s) and subject(s) for which data were collected and the numbers of schools, teachers, and students participating at each site during year 1.

DATA COLLECTION

We collected three types of data at each site: student achievement test scores, teacher questionnaire responses, and student demographics. Data were collected in the spring of 1997.

Table 2.1

Details of Year 1 Participating Sites

Site	Grade	Subject	Number of Schools[a]	Number of Teachers	Number of Students (varied by test in some sites)
1	3rd	Math	17	46	804
2	5th	Math	20	100	1,651–1,686
2	5th	Science	20	99	1,639–1,662
3	5th	Math	18	73	1,366–1,451
3	5th	Science	20	74	1,367–1,438
4	5th	Science	19	45	909–932
5	7th	Math	17	48	2,937–3,018
5	7th	Science	19	33	2,047–2,079
6	7th	Math	25	57	3,237
6	7th	Science	25	52	3,279

[a]Some schools at each site are included in both the mathematics and science samples.

Student Achievement Data

We obtained student scores on the mathematics and science assessments regularly administered at each site and supplemented these with additional assessments, where feasible, to provide both multiple-choice and open-response scores. Supplementary tests were chosen in consultation with local staff, who were encouraged to select measures they believed were reasonably well aligned with their reform efforts. Hands-on science tasks developed by RAND were made available, and some sites opted to use them. Mosaic project staff trained exercise administrators to administer some of the supplementary measures, including RAND's hands-on tasks. All other tests were administered by the classroom teachers or by test administrators who worked at the local sites. Table 2.2 shows the types of tests administered at each site. Wherever possible, we used existing tests, including state-developed tests and commercially available standardized tests. The column headed *Added for Mosaic Study?* in-

Table 2.2

Sites, Subjects, and Assessment Instruments

Site	Grade	Subject(s)	Tests	Format[a]	Added for Mosaic Study?
1	3[b]	Math	State[c]	MC	No
			State	OR	No
			State	Grid-in	No
2	5	Math	State[c]	MC	No
			Stanford 9[d]	OR	Yes
		Science	Stanford 9[d]	MC	Yes
			RAND Levers and Friction[e]	OR (hands-on)	Yes
3	5	Math	CTBS[f]	MC	No
			Stanford 9[d]	OR	Yes
		Science	CSIAC[g]	MC	No
			CSIAC[g]	OR (hands-on)	No
4	5	Science	CSIAC[g]	MC	No
			CSIAC[g]	OR (hands-on)	No
5	7	Math	State[c]	MC	No
			Stanford 9[d]	OR	Yes
		Science	Stanford 9[d]	MC	Yes
			RAND Levers and Classification[e]	OR (hands-on)	Yes
6	7	Math	MAT7[h]	MC	No
		Science	MAT7[h]	MC	No

[a]MC = multiple-choice; OR = open-response.

[b]At this site, we studied teaching practices for third-grade teachers and measured the relationships with student test scores gathered during the following fall, when students had advanced to the fourth grade.

[c]Refers to tests developed by the state.

[d]Stanford Achievement Test Series, Ninth Edition, published by Harcourt-Brace Educational Measurement.

[e]See Stecher and Klein, 1996, for a description of tasks and scoring guides.

[f]Comprehensive Test of Basic Skills, published by CTB/McGraw-Hill.

[g]California Systemic Initiatives Assessment Collaborative. This test was developed by a consortium of educators and researchers and was designed to be aligned with NSF-supported reform efforts.

[h]Metropolitan Achievement Tests, Seventh Edition, published by Harcourt-Brace Educational Measurement. At this site, we were unable to schedule any open-response testing.

dicates whether we supplemented the district's or state's testing program with additional measures or relied only on those measures already used by the sites.

Teacher Questionnaires

Our primary measure of teaching practices at each site was a modified version of a questionnaire developed and used extensively by Horizon Research, Inc. (HRI) to evaluate the implementation of the Local Systemic Change (LSC) initiatives. Questionnaires were administered to all teachers in each school teaching the targeted subject and grade level. Typically, the site coordinator or assistant distributed the questionnaires either individually or at after-school meetings and then collected completed questionnaires in individual sealed envelopes for return to RAND.

We created separate questionnaires for mathematics and science teachers, but many of the items were identical across subjects. Teachers were asked to report the frequency of various instructional practices ranging from traditional (e.g., "Have students watch me [teacher] do a science demonstration") to reform ("[Students] conduct investigations where they develop their own procedures for addressing a question or problem"). General topics included the amount of time spent on science/mathematics, approach to introducing a new topic, typical teacher instructional practices, typical student activities, types of written assignments, teachers' use of students' written work, and methods of assessing student learning.

Although NSF did not mandate a particular curriculum or a specific set of teaching strategies for the SIs, there was an emerging consensus among mathematics and science educators about what should be taught and how it should be presented (National Council of Teachers of Mathematics, 1989; National Research Council, 1996). In light of this consensus, it is not surprising that the systemic reform programs adopted very similar content and instructional goals. An independent evaluation of the SSIs reported that "across the states there was remarkable similarity in the perceived shortcomings of current practices and the set of desirable reforms in curriculum content and instructional strategies" (Shields, Marsh, and Adelman, 1998, p. 2). The shared content goals included greater emphasis on

the understanding of mathematics and science concepts, the application of this knowledge to everyday situations, and the integration of concepts across subjects. The instructional emphasis was equally distinct. The reforms sought to have instructors engage students actively in their own learning, to be sensitive to each student's learning style, to increase the use of technology, and to utilize new forms of assessment for instructional planning, rather than viewing students as passive learners who absorb unrelated facts and procedures. In mathematics, this meant more "data gathering and analysis, statistics, geometry and visualization, discovery learning, and 'constructivist' approaches"; in science, more "scientific processes, such as observation, comparison, experimentation, hypothesis generation, hypothesis-testing, and theory building" (New Jersey SSI Proposal, 1992, p. 7; quoted in Shields, Marsh, and Adelman, 1998, p. 3). Our measures of instructional strategies were designed to be consistent with this espoused commonality of purpose.

In addition, each teacher was asked to complete a brief demographic section, providing information about his or her college degree, teaching certification, coursework in mathematics and/or science, gender, ethnicity, and years of teaching experience. At sites where instruction was delivered by science or mathematics specialists instead of the regular classroom teacher, we administered surveys to the specialists and also asked the respondents to clarify their teaching situations.

HRI, acting as a subcontractor to RAND, developed the questionnaire, processed the data, and prepared analysis files. HRI also validated the instruments at one site in which RAND selected a sample of schools whose teachers were expected to have a wide range of implementation of the reforms. A local coordinator scheduled interviews with 40 teachers from these schools. Trained staff from HRI interviewed each teacher about a recent science or mathematics unit; the teachers were asked to discuss their goals for the unit, the extent to which students engaged in investigations and collaborations, and the types of assessment they used. Teachers were also asked to show associated artifacts (student work, journals, assignments, etc.). In some cases, the validator observed an actual lesson.

On the basis of the interviews, artifacts, and observations, the validator rated the instructional program of each teacher on a five-point

scale on each of three dimensions: the use of student-centered strategies, the investigative culture of the classroom, and the use of reform-oriented strategies. The validator also made an overall judgment of the degree to which each teacher's instructional program reflected reform practice. These ratings were compared with the self-reported practice information from the teacher surveys. Reform-practices scales for science and mathematics were created by combining teachers' responses to items on the surveys that reflected standards-based instruction in each subject. We identified the items for these scales using factor analytic techniques and the judgment of recognized experts in science and mathematics instruction. A combined scale was computed by summing the results from the mathematics and science scales.

There was a statistically significant positive relationship between the validators' overall ratings and the combined survey scales: The Spearman-Rho correlation coefficient between these measures was 0.44. It is difficult to say what degree of correspondence we would expect to find between these two distinct measures of practice. The interviews and observations focus on one particular lesson, and they are probably sensitive in unknown ways to the dynamics of the interaction between teacher and interviewer. The surveys emphasize a longer span of time and a wider range of content, but they are subject to unknown self-report biases. In addition, the validators' ratings included their impressions of the quality of the instruction, whereas the questionnaires addressed only the frequency with which specific reform practices were used. Finally, the validators exhibited a tendency to assign scores near the middle of the five-point scale, resulting in low variability and restricted range among the ratings. In view of these considerations, an overall correspondence of 0.44 is reasonable.

Demographic Data

Information about student characteristics was obtained from the sites for three purposes: (1) to verify that comparison schools were similar to implementing schools in terms of student demographics, enrollment, and grade span; (2) to be included as covariates in the analysis of relationships between teaching practices and student

achievement; and (3) to enable us to study whether these relationships varied as a function of student characteristics.

At most sites, we obtained data on students' race/ethnicity, gender, participation in free or reduced-price lunch programs, language background, participation in special education or gifted programs, and test scores from the previous year. We did not obtain the same set of covariates from all sites, partly because some of the covariates did not apply to particular student populations. For example, some sites have large numbers of students with limited English proficiency (LEP), whereas others have none; some sites exclude special education students from testing, and some include them. A few of the covariates, such as age and participation in a gifted program, were unavailable from some sites. Excluding these from the models had virtually no effect on relationships between student achievement and teaching practices.

PARTICIPATION RATES

Rates of participation in the study by schools, teachers, and students were quite high. It is difficult to summarize participation in a simple manner because of the multistage nomination and enrollment process, differences in procedures across sites, and partial participation by some schools, teachers, and students. Nevertheless, across the six sites, between 85 and 100 percent of the schools we initially contacted participated in the study, i.e., tests and teacher surveys were administered and completed by the majority of eligible individuals. Similarly, at the site level, between 71 and 98 percent of the teachers who received a survey completed it. Student participation rates were more difficult to compute because of partial participation and because of a variety of testing exclusions. In addition, we could not always link students and teachers. Nevertheless, across the six sites, between 65 and 94 percent of the students identified as being taught by the teachers in the study completed all the desired tests in mathematics or science. Detailed descriptions of school, teacher, and student participation at each of the sites are given in Appendix A.

ANALYSIS

We investigated the degree to which student achievement was associated with teachers' use of instructional practices consistent with

the reforms, using linear regression analysis, which enabled us to control for student background characteristics and previous test scores. We found that teacher background variables did not provide any additional explanatory power, and therefore we do not include them in the results reported here. At each site, we conducted separate analyses for mathematics and science, for open-response and multiple-choice tests, and for reform and traditional practices. We fit these models using individual student data, with all students from the same classroom receiving the same values for the reform and traditional scales, and we used an adjusted standard-error estimate to account for possible correlation among responses from students with the same teacher (McCaffrey and Bell, 1997).

In addition, the use of data from multiple sites provided an opportunity to conduct a planned meta-analysis. We therefore also conducted pooled analyses, which combined data from all six sites to produce a single estimate of the coefficient relating teaching practices (reform or traditional) to student achievement. We conducted separate analyses by subject (mathematics or science), test format (multiple-choice or open-response), and teaching-practices scale (reform or traditional), for a total of eight pooled analyses.

Pooled analyses are appropriate when the coefficients from the various sites describe a single relationship, as was the case in this study. We examined the relationship between teaching practices and student test scores at all six sites, using the same study design and similar analyses. Specifically, we assumed that the coefficients from our models are homogeneous and that small differences in our site studies (e.g., differences in tests and the covariates in our models) can be modeled as small random variations among the coefficients. The homogeneity assumption was tested by examining the variability among the coefficients across the sites. Technical details on the pooled analyses and the estimates of variability are given in Appendix D.

For the pooled analyses, we used the estimated regression coefficients described earlier. We did not pool the individual student data, because we used different covariates in our site models, and thus we would have had to exclude some useful predictors from the model. However, the estimates we obtained by pooling the regression coefficients for instructional practices are similar to those we would

have obtained by pooling individual scores and fitting a random-coefficients model with interactions between sites and the covariates (Goldstein, 1995). Thus, our approach enabled us to pool data across sites without requiring identical models for every site.

Chapter Three

FIRST-YEAR RESULTS

In this chapter, we present summaries of teachers' reported use of reform and traditional practices and our findings with regard to the relationships between use of these practices and student achievement at each site. Finally, we describe the results of an analysis of differences between open-response and multiple-choice achievement measures.

DISTRIBUTIONS OF TEACHING PRACTICES

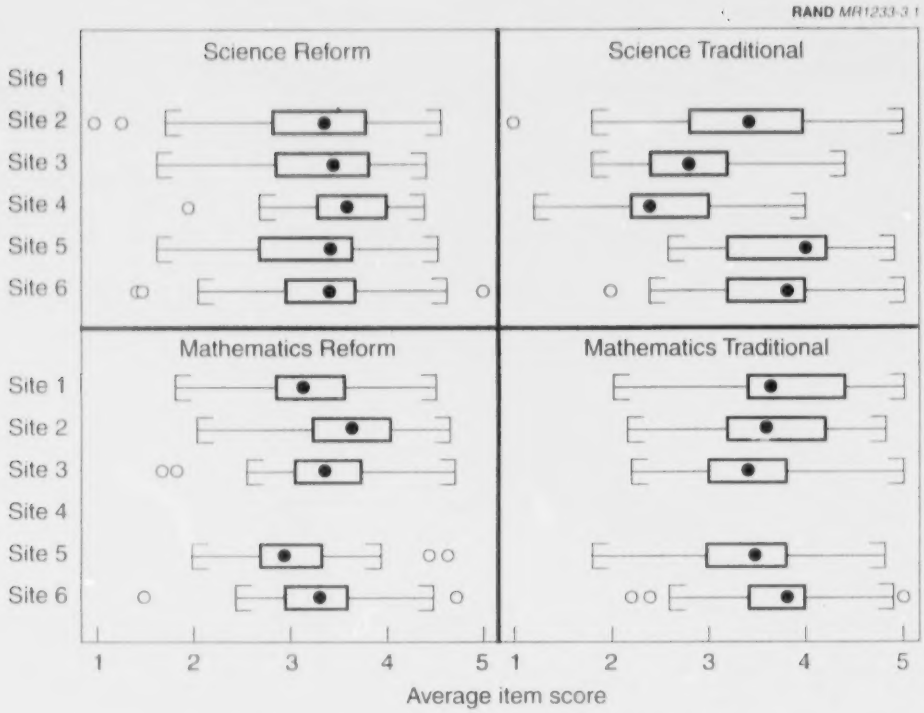
On the basis of exploratory factor analyses of the questionnaire items, we identified two clusters of items and created scales from them by simply summing the scores on each item. The first scale measured the teachers' use of reform practices. Teachers were asked to report the frequency of use of 22 specific reform practices (e.g., cooperative groups, portfolios, and extended investigations). We also created a five-item traditional-practices scale based on items that measured the amount of time teachers spent on traditional teaching practices (e.g., textbook work, lectures, and short-answer tests). Appendix B lists the items in each scale. The distinction between reform-related practices and traditional practices that emerged from factor analyses conducted on each site's data is consistent with the kinds of definitions used in other research on mathematics and science instruction reform (e.g., Cohen and Hill, 1998). However, it is important to note that the two scales are not opposites of one another. A principal-components analysis of the questionnaire data identified these two separate scales at each site, suggesting that teachers may use both reform and traditional practices to different

degrees. Correlations between the two scales ranged across sites from moderately negative to moderately positive, with many close to zero. It is possible for teachers to be high on both scales, because the scale scores do not indicate the total amount of time spent on these practices, but rather the frequency with which they are used. Thus, a teacher who intersperses lecture-style teaching with opportunities for student discussion in every lesson might score high on both scales. In addition, there are other activities not addressed by either scale, so it is possible for teachers to receive low scores on both.

At each site we found a wide range of practices on both the reform and the traditional scales. The box-and-whiskers plots in Figure 3.1 show the distributions of mean scale scores for each combination of site and subject (mathematics or science). For these plots, the score for each teacher was simply the average item response across items. All items used a five-point Likert scale, so teachers' scores could range from 1 (rarely or never using any of the practices) to 5 (engaging in all practices daily or almost daily). The solid dots indicate the average score for all teachers. The lower end of the box is the 25th percentile of the distribution, and the upper end of the box is the 75th percentile; the whiskers show the extreme points, excluding outliers. Outlier values are shown as individual points in the plots.

Overall, the scores were similarly distributed across sites. Science teachers' average scores on the reform scale ranged from 3.27 at site 2 to 3.57 at site 4 (upper left quadrant of the figure). Site averages on the reform scale for mathematics teachers were somewhat more variable, ranging from 3.01 to 3.61 (lower left quadrant). On the traditional-practices scale, site averages for science were more variable than those for mathematics. The former ranged from 2.65 to 3.78 (upper right quadrant), whereas the latter ranged from 3.33 to 3.73 (lower right quadrant). Interestingly, sites at which teachers' average use of reform practices was quite similar showed fairly large discrepancies on the traditional-practices scale.

We also found substantial variation in teaching practices within schools (not shown in Figure 3.1), regardless of the degree of participation in the reform programs. Clearly, some teachers in participating schools had not adopted many of the reform practices emphasized by the SI, whereas some teachers in nonparticipating schools



NOTE: Box-and-whiskers plots show mean score (dot), 25th to 75th percentile range (rectangle, or box), and extreme points (whiskers). Outliers appear as open circles.

Figure 3.1—Distribution of Teacher Scores on Reform and Traditional Scales by Site and Subject

were using these practices even though they had not been exposed to SI-specific professional development. This underscores the importance of using classroom-level measures of teaching practices rather than studying differences at the school level.

RELATIONSHIPS BETWEEN TEACHING PRACTICES AND STUDENT ACHIEVEMENT

Our analyses relating teaching practices to student achievement showed that teachers' use of reform practices appeared to be positively related to student achievement at most sites, but the effects were quite small and rarely reached statistical significance. Use of traditional practices, by contrast, was often negatively related to stu-

dent achievement, particularly in mathematics, but again the relationships were weak.

Figures 3.2 through 3.9 provide an overview of our findings in the six year 1 sites. The full regression models are given in Appendix C. The relationships reported in the figures are the estimated coefficients from our regression models for the reform- and traditional-practices scales. We standardized test scores and teaching-practices scales so that the reported coefficient is the expected difference in test score standard-deviation units for a one-standard-deviation unit increase in scores on the reform or traditional scale. The dark dot represents the point estimate for the coefficient, and the gray bar represents the 95 percent confidence interval for that point estimate. When the bar does not meet the zero line, the coefficient is statistically different from zero. The lowest bar in each figure shows the average coefficient from the pooled analysis, described later.

Figure 3.2 shows relationships between the use of reform practices and achievement on open-response mathematics tests. At four of the five sites that had open-response mathematics tests, higher scores were associated with greater use of reform practices. However, the coefficients were statistically significantly greater than zero at only two of these sites. Similarly, Figure 3.3 shows that for almost all of the participating sites, higher multiple-choice test scores in mathematics were associated with greater use of reform practices, but none of the estimates was statistically significantly different from zero. Figures 3.4 and 3.5 show that greater use of reform practices in science was associated with higher test scores on both open-response and multiple-choice measures. Again, most of the estimated coefficients were extremely small and were not statistically significantly different from zero, even though coefficients across sites show a consistent pattern of a weak positive relationship between the reform-practices scale and test scores.

In contrast, the majority of the relationships between the use of traditional practices and student achievement were negative. For example, Figure 3.6 indicates that at all of the participating sites, greater use of traditional teaching practices in mathematics was associated with lower scores on open-response mathematics tests. However, none of the estimated coefficients for traditional practices was statistically significantly different from zero.

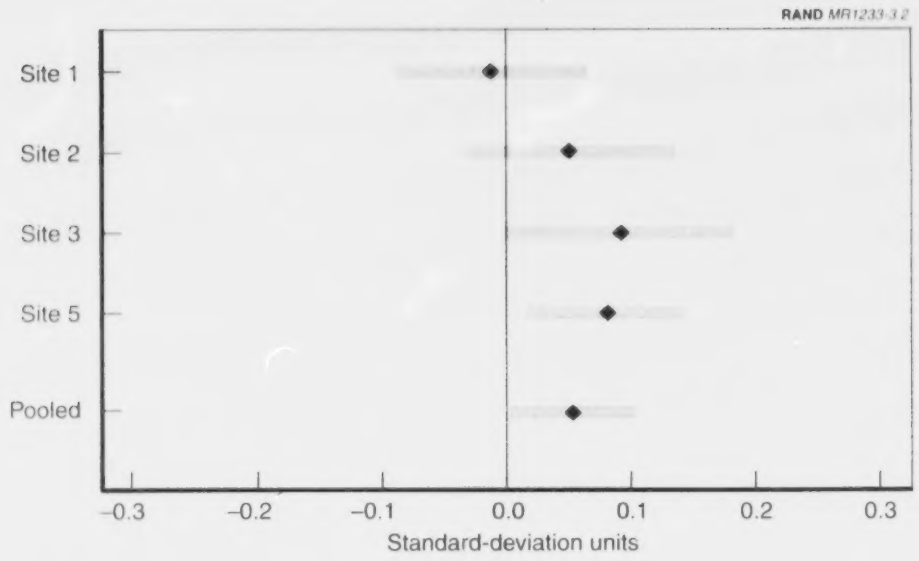


Figure 3.2—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Open-Ended Tests

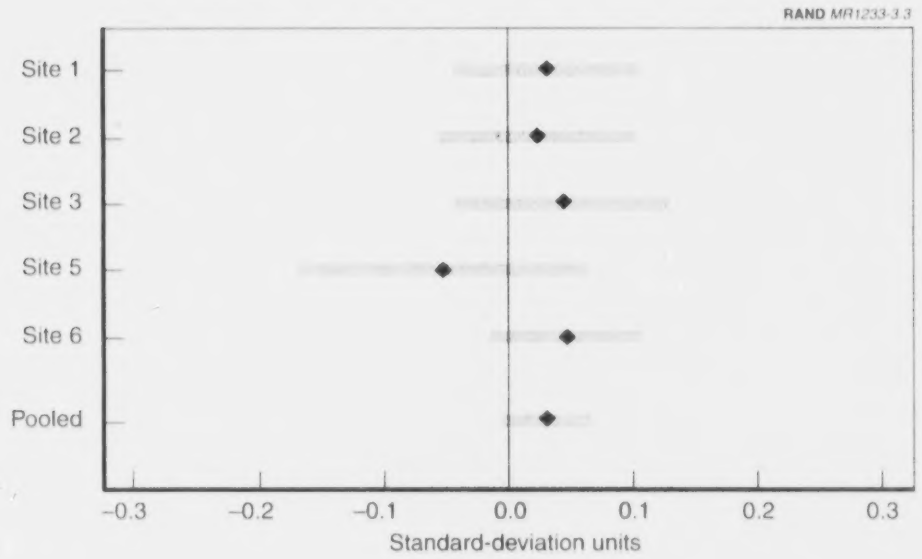


Figure 3.3—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Multiple-Choice Tests

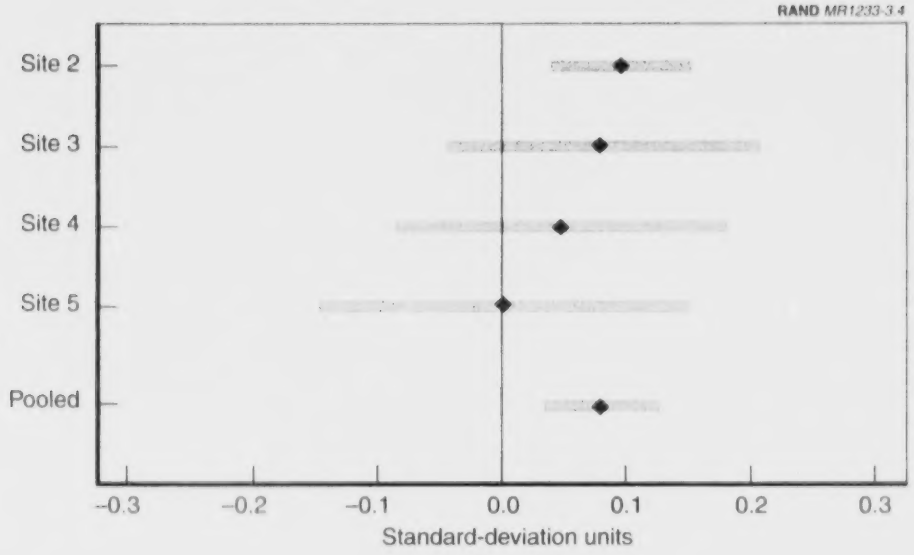


Figure 3.4—Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Open-Ended Tests

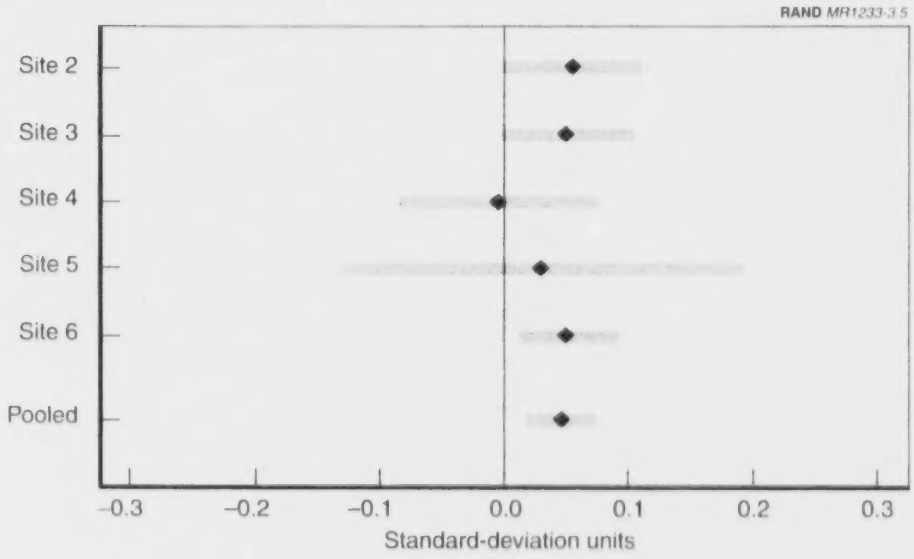


Figure 3.5—Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Multiple-Choice Tests

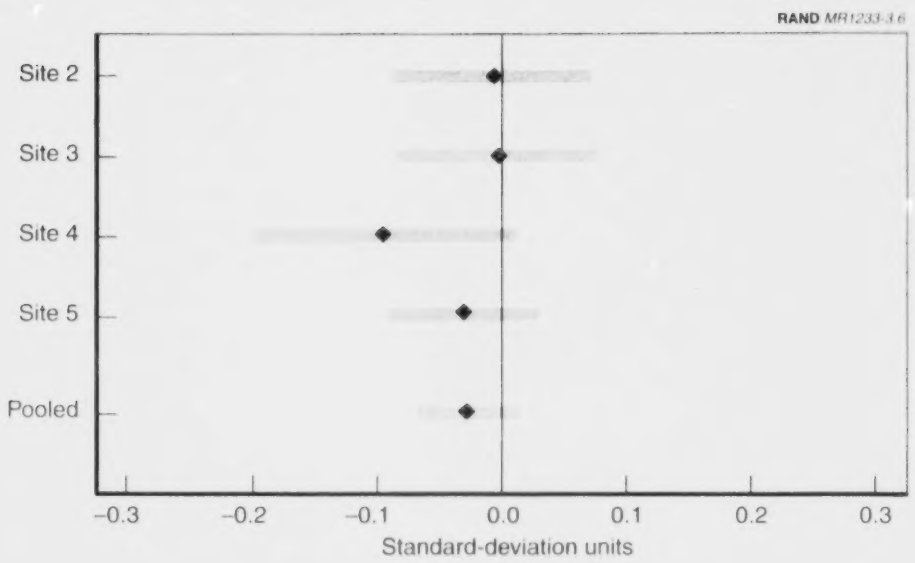


Figure 3.6—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Open-Ended Tests

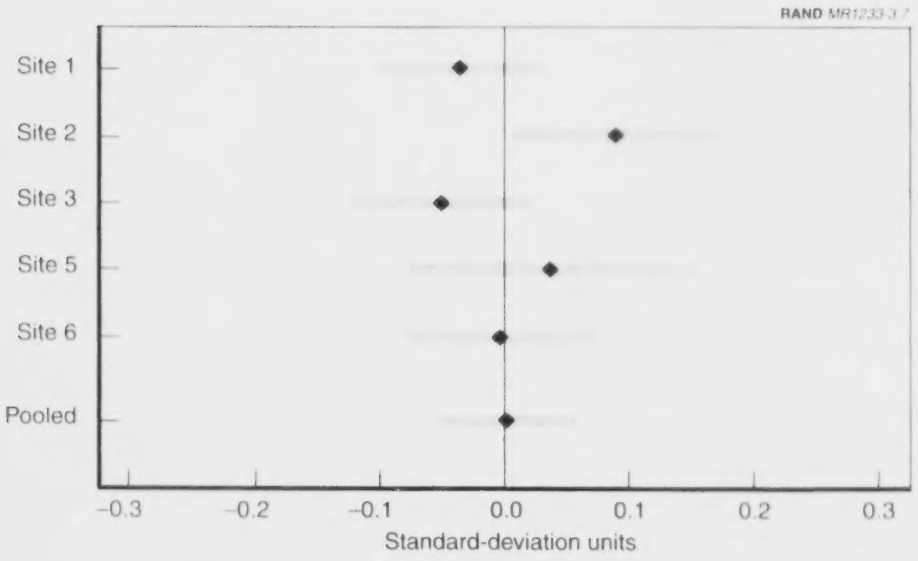


Figure 3.7—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Multiple-Choice Tests

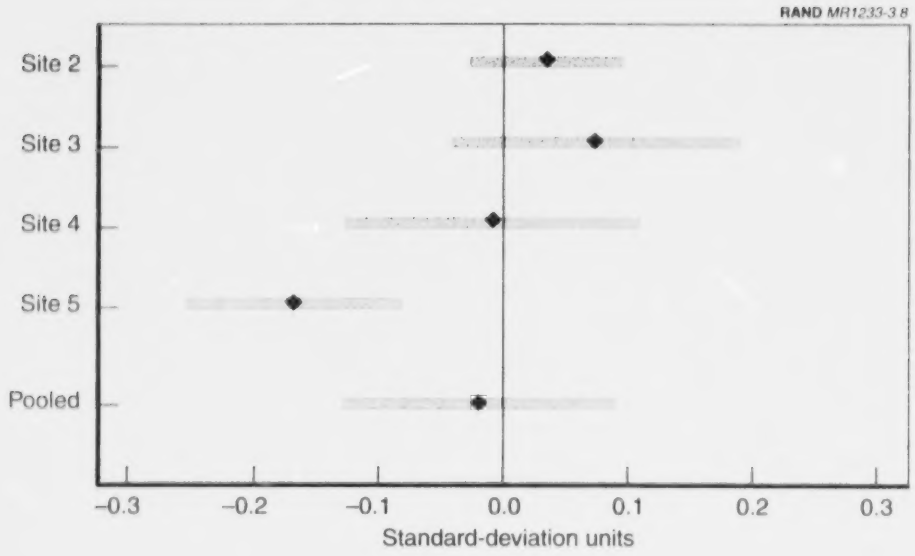


Figure 3.8—Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Open-Ended Tests

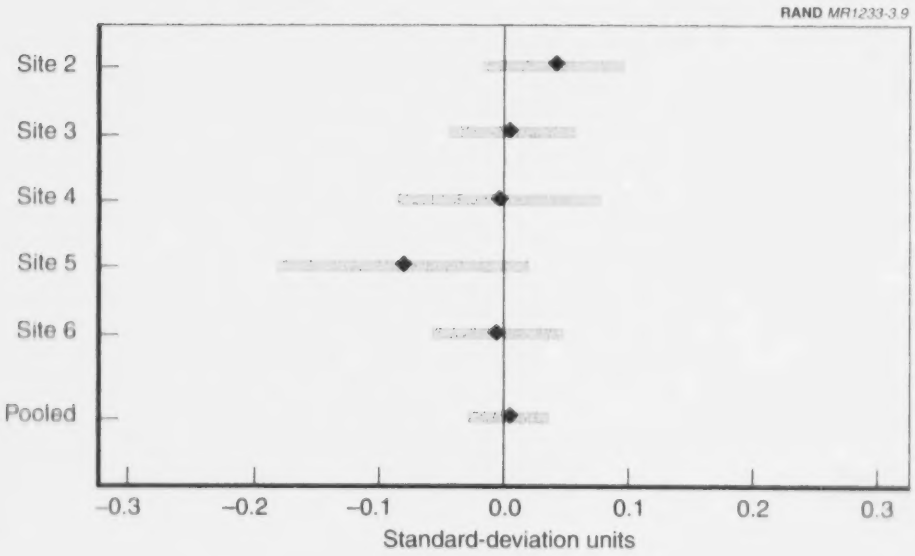


Figure 3.9—Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Multiple-Choice Tests

The relationship between teaching practices and test scores is at most small in almost all our models. For example, the largest positive relationship was found between reform teaching practices and open-response science tests at site 2 (see Figure 3.4), where the standardized regression coefficient was 0.09. Our model suggests that with a teacher at this site using all of the reform practices monthly, the average student was predicted to score at about the 48th percentile on the test, while for a teacher using all of the reform practices weekly, we would predict that a similar student would score at about the 54th percentile on the test.[1] Smaller changes in percentiles would be expected at the other sites. Compared with the coefficients for most of the student background characteristics (e.g., an average coefficient of 0.54 across sites for participation in free and reduced-price lunch programs), all of the relationships we observed may be considered small.

We expected to see larger relationships between reform practices and open-response measures than between reform practices and multiple-choice measures because the former tend to be more closely aligned with the reforms. Inspection of the regression coefficients suggests that this is the case. Later we discuss a test of the statistical significance of this difference.

The bottom bars in Figures 3.2 through 3.9 show the pooled estimates of the standardized regression coefficients for each of the eight analyses. The coefficients and confidence-interval bounds are presented in Table 3.1, and additional detail is provided in Table D.1 in Appendix D. In most of the analyses, the variability of coefficients across sites was sufficiently small to be within the range expected as a result of sampling error within sites. In these cases, the pooled analysis is appropriate. In analyses where we did find variability (discussed below), the pooled estimate is difficult to interpret because it represents an average over a set of disparate coefficients.

---

[1]We used our model to predict the score for the "average" student (a student with all student background predictors set to the mean) with a teacher scoring 3 on each reform-practices item (monthly use of reform practices). We then found the percentile of this predicted score among the test scores from the site and repeated the process for the average student with a teacher scoring 4 on each item (daily use). The percentile is based on our sample and is not a percentile from a national norming group.

Table 3.1

Coefficients from Pooled Analyses of Relationships Between Instructional Practices and Achievement

Subject	Test Format	Scale	Weighted Average Coefficient	Lower Bound of 95% Confidence Interval	Upper Bound of 95% Confidence Interval
Math	OR	Reform	0.053	0.008	0.098
Math	MC	Reform	0.030	–0.003	0.063
Science	OR	Reform	0.079	0.036	0.122
Science	MC	Reform	0.045	0.022	0.069
Math	OR	Traditional	–0.025	–0.061	0.012
Math	MC	Traditional	0.001	–0.049	0.052
Science	OR	Traditional	–0.018	–0.123	0.088
Science	MC	Traditional	0.006	–0.023	0.034

Note: OR = open-response; MC = multiple-choice.

One instance in which the variability of coefficients across sites was greater than zero was the relationship between reform practices and student achievement on open-response mathematics tests (Figure 3.2). The pooled coefficient was 0.053, but the test of variability indicated heterogeneity across sites. The variation was primarily a result of the negative relationship (–0.010) we observed for site 1; the other three coefficients ranged from 0.052 to 0.092. When we conducted the pooled analysis excluding site 1, the estimate was 0.075, with no indication of heterogeneity in the coefficients. This difference is probably due to the fact that at site 1 we used scores on a locally developed open-response mathematics test, whereas at the other three sites we administered the Stanford 9 test. The relationship between instructional practices and achievement may be sensitive to the particular instrument used or to the administration conditions. This problem is discussed further in Chapter Four.

Our pooled estimate of the relationship between reform teaching practices and student achievement on multiple-choice mathematics tests was 0.03, not statistically significantly different from zero. For traditional teaching practices, we obtained pooled estimates that were slightly less than zero for both multiple-choice and open-response tests in mathematics, but neither was significantly different from zero. The variability among coefficients was sufficiently small

to permit pooling for all of these analyses except for that of traditional practices and multiple-choice tests; but again, this variability resulted primarily from a single outlier, the estimate of 0.09 from site 2. Although the pooled estimates for reform and traditional teaching practices are of different sign, the uncertainty in each estimate is substantial enough that the confidence intervals for the two estimates overlap. Thus we cannot rule out the possibility that the observed differences are only a result of sampling error.

We obtained pooled estimates of 0.045 for the relationship between reform teaching practices and multiple-choice test scores in science, and 0.079 for the relationship between reform teaching practices and open-response test scores. Both of these estimates were statistically significantly different from zero.

For traditional teaching practices in science, we obtained estimates of –0.018 for open-response scores and 0.005 for multiple-choice scores, neither of which differed significantly from zero. Both analyses revealed variability among coefficients relating science scores to traditional practices, suggesting that a pooled analysis might not be appropriate in these cases. Again, there is substantial uncertainty in the pooled estimates for both the traditional and reform scales for both types of science tests, and we cannot rule out the possibility that the observed differences are simply a result of sampling error.

To summarize, the pooled analyses revealed statistically significant positive relationships between teachers' use of reform practices and achievement on both kinds of tests and in both subjects. However, these relationships are much smaller than the relationships between test scores and other covariates, such as ethnicity and socioeconomic status (see Appendix C). Results for open-response mathematics scores appear somewhat sensitive to the particular test used, but for the other measures we detected no evidence of heterogeneity among sites in the strength of the relationships between reform practices and achievement.

In general, teachers' use of traditional practices was unrelated to student achievement. However, our measure of traditional practices is less reliable than our measure of reform practices. Across sites and subjects, the average alpha coefficient is 0.70 for traditional practices, while that for the reform-practices scale is 0.92. The lower reli-

ability would tend to attenuate the relationship between the traditional practices and test scores and might contribute to the weakness of the estimated relationship between traditional practices and outcomes.

ALTERNATIVE FORMULATIONS FOR SITE MODELS

There are several alternative approaches we could have taken to model the relationship between instructional practices and student achievement. Because prior-year test scores were unavailable at sites 1 and 5, we used contemporaneous scores, but we could instead have omitted the achievement covariate altogether. We also chose to explore the reform and traditional scales in separate models, but we could have put them together in the same model. To explore the effects of our modeling decisions, we conducted some analyses using alternative model specifications. Appendix F discusses the results of our explorations of contemporaneous test scores, and Appendix G presents results of our analyses of the effects of combining the reform and traditional scales in a single model. The analyses revealed that including contemporaneous scores probably resulted in conservative estimates of the coefficients for instructional practices, but the effect is small. Including reform and traditional scales in the same model likewise had little effect on our results.

DIFFERENCES BETWEEN TEST FORMATS

Consistent with the individual site results, inspection of the coefficients from the pooled analyses suggested slightly larger relationships between open-response scores and reform teaching practices than between multiple-choice scores and reform teaching practices. This finding is consistent with the hypothesis that open-response tests tend to be more closely aligned with the reforms and therefore better able to indicate effects. To test the statistical significance of this difference, we calculated the difference in standard-deviation units between each student's score on the open-response test and his or her score on the multiple-choice test in the same subject. We then modeled these differences as a function of teaching practices and student background covariates. The analysis was repeated for both subjects and for all sites.

The coefficients for teaching practices obtained for each site are presented in Figures 3.10 through 3.13 and in Table E.1 in Appendix E. The coefficient for reform teaching practices was positive for three of the four sites where we collected data on mathematics achievement. However, only one of the differences, 0.113 for site 5, was statistically significant. Similarly, the coefficient for reform teaching practices was positive for three of the four sites where we collected data on science achievement, but none of these differences was statistically significant.

We again conducted a pooled analysis of coefficients across sites. Results are given in Table E.2 in Appendix E. For mathematics and the reform scale, the pooled estimate was 0.032. This implies that across the sites, the expected increase in student mathematics test scores for a unit increase in a teacher's score on the reform scale was 0.032 standard-deviation units higher for open-response tests than for multiple-choice tests. However, our estimate was not statistically significantly different from zero. In addition, we found a relatively large between-site variance in these estimated differences, even after controlling for sampling error within sites. In other words, we found that the difference in the sensitivity of open-response and multiple-choice tests varied from site to site. At the two fifth-grade sites where the open-response test was the Stanford 9, the differences were similar, 0.032 and 0.051. At site 5, we again administered the Stanford 9 open-response test, but this time to seventh graders, and the difference between open-response and multiple-choice tests was 0.113. At site 1, we used scores from a test developed by the state, and the difference was –0.041. Hence, the sensitivity of tests might depend on both the test form and the grade. Additional data are necessary to explore this hypothesis.

For science and the reform scale, the pooled estimate was 0.031. This implies that across the sites, the expected increase in student science test scores for a unit increase in a teacher's score on the reform scale was 0.031 standard-deviation units higher for open-response tests than for multiple-choice tests. Again, our estimate was not statistically significantly different from zero. For science, there was little variability in estimated differences across sites after we accounted for sampling error within sites.

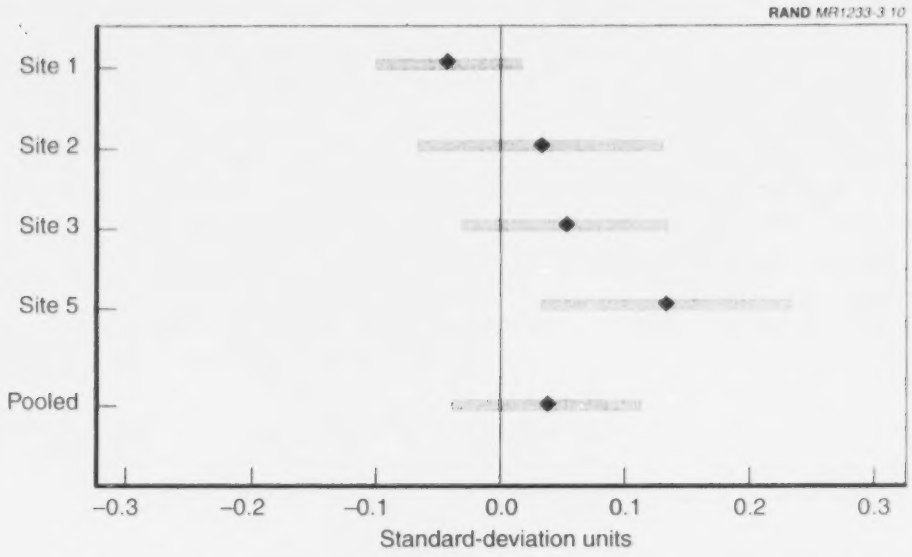


Figure 3.10—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Difference Between Open-Ended and Multiple-Choice Tests

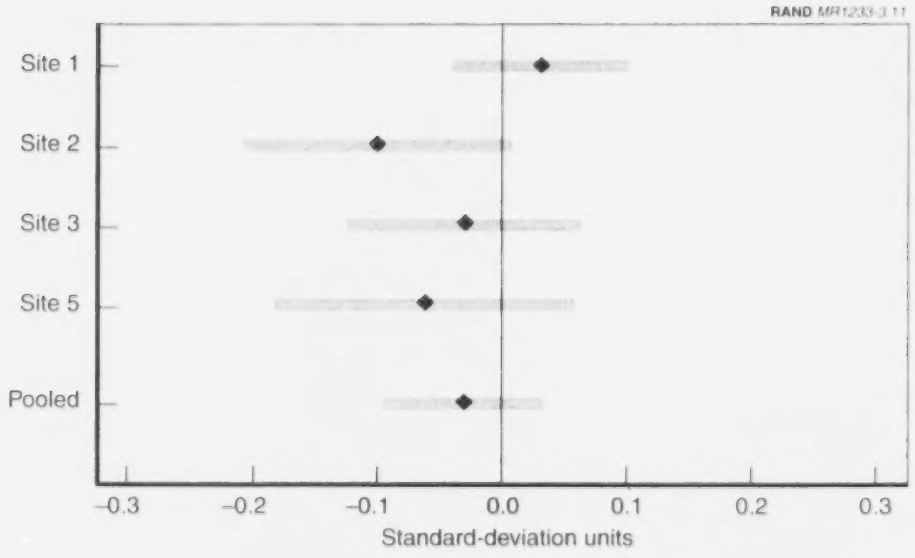


Figure 3.11—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Difference Between Open-Ended and Multiple-Choice Tests

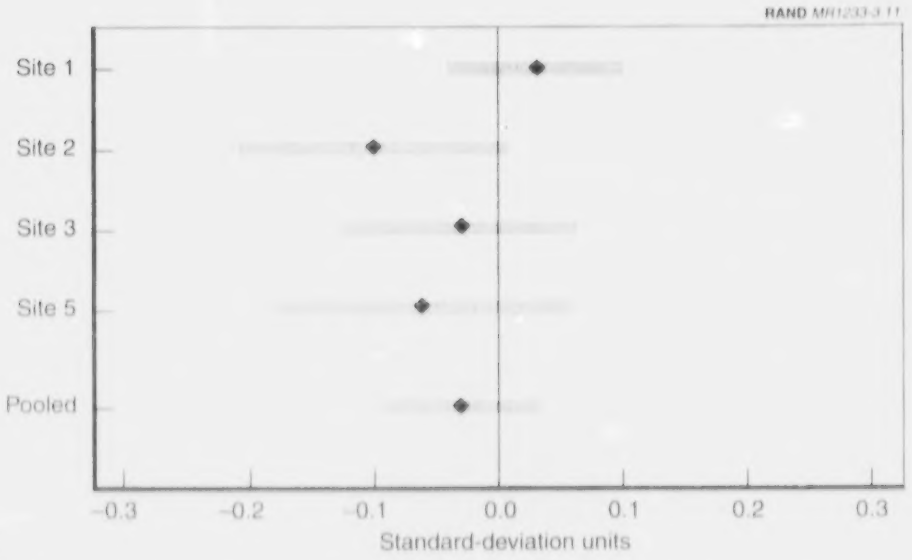


Figure 3.12—Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Difference Between Open-Ended and Multiple-Choice Tests

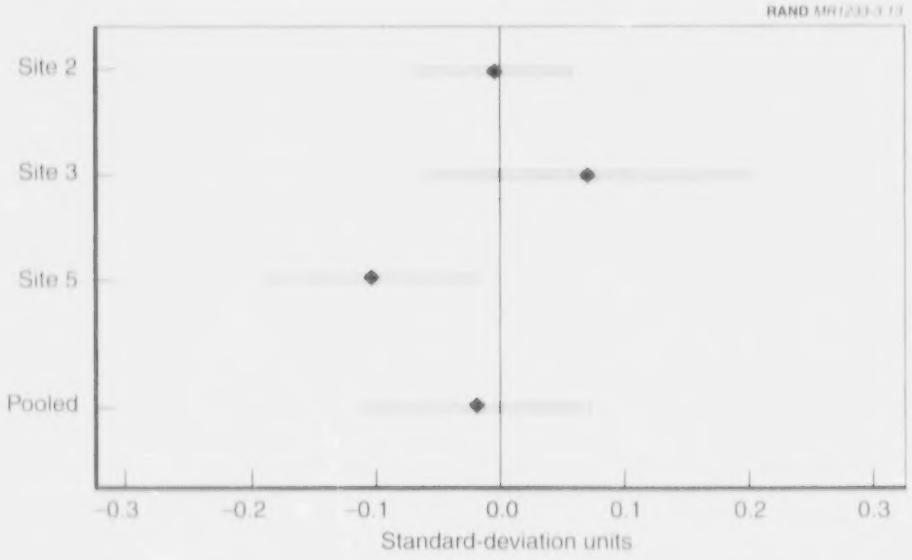


Figure 3.13—Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Difference Between Open-Ended and Multiple-Choice Tests

Thus, although inspection of regression coefficients suggests that open-response tests function differently from multiple-choice tests, our data do not provide sufficient evidence to support the claim that the formats differ in their sensitivity to the effects of the reform. Even so, the consistency in the patterns we observed and the fact that educators involved in these reforms believe that open-response tests are generally more closely aligned with their efforts suggest that further investigation is needed to explore format differences. As states continue to develop standards-based assessments, and as results from these assessments are increasingly used in evaluations of educational programs, it is critical that the validity of different test formats be examined.

STUDY LIMITATIONS

Several caveats must be considered when interpreting the results of this study. As in most educational research, our inability to investigate effects by using an experimental design places some limitations on the kinds of inferences that can be made from results. Perhaps the primary problem is that without random assignment of students and teachers to treatments, we cannot be certain that the relationships we observed can be attributed solely to classroom practices. There may be other differences in student characteristics across classrooms that contribute to differences in performance and that influence what teachers do. For example, teachers may tend to engage in more reform-based practices with higher-achieving students. Controlling for prior achievement is helpful, but it does not eliminate the problem completely.

A second limitation is the lack of information on what led teachers to use particular practices. Some may have adopted certain strategies as a result of participation in the professional development activities that are provided by the SI funds, but there are many other potential sources. The large variability in teaching practices within schools, which was observed for SI as well as non-SI schools, suggests that factors other than SI participation are influencing teachers' decisions about how to teach. Determining the reasons for teachers' use of practices was not the initial intent of our study, but information on those reasons would be helpful to people who are designing and implementing professional development programs.

A third weakness of our approach stems from the use of questionnaires to measure instructional practices. As on any questionnaire, our items are subject to inaccurate responses, particularly those items that reflect social desirability. Perhaps more important, our questions addressed only the frequency with which teachers used particular practices and did not address the way in which they were used or the overall quality of the instruction. Clearly, some approaches to using cooperative groups are more effective and more consistent with the intent of the reform than others are, but we cannot detect these differences on the basis of our questionnaires. Multiple classroom observations, interviews, and inspection of classroom materials would undoubtedly provide a better measure of instructional practice. This type of data, however, is considerably more expensive to collect and is usually collected only on a small scale.

Finally, as discussed earlier, our analyses focused on students' exposure to practices during a single academic year, and therefore we were not able to follow the development of teacher experience in reform practices or the impact of student exposure to these practices across several years. Many of the achievement tests used at our sites require students to apply knowledge and skills they have gained over a number of years, so performance on these tests is undoubtedly influenced by students' instructional experiences in prior years. Project directors at some of our sites reported that student exposure to reform practices is typically small, even at sites that have been grant recipients for several years, because programs generally need time to be fully implemented. It is widely believed that students must be exposed to reform practices for more than a single year before the effects of these practices on achievement can become clearly evident, but little information is currently available to support or refute this claim.

Chapter Four

DISCUSSION

In this chapter, we briefly summarize our findings from year 1 of the Mosaic study and discuss their implications. We also suggest possible explanations for the variability in findings across sites, and we then describe plans for future data collection and analysis.

SUMMARY OF YEAR 1 FINDINGS

As illustrated by Figures 3.2 through 3.9, the relationships between student achievement and teachers' use of instructional practices supported by the SI reforms tend to be positive but small, particularly in comparison with relationships between achievement and student background characteristics such as socioeconomic status and ethnicity. If, in fact, the observed relationships represent the effects of teaching practices on student achievement, their small magnitude may not be surprising, given the brief period of time (less than one academic year) captured by teachers' questionnaire responses. Use of particular instructional strategies in a single course during a single school year would not be expected to lead to effects as large as those associated with student background characteristics. Several years of exposure to instructional reforms may be needed to achieve a reasonably large effect. This suggests the need for longitudinal investigations, discussed below.

The direction of relationships was fairly consistent across sites, but there was some variation in magnitude. This variation may come from several potential sources. First, our models differed slightly across sites because we relied on locally available data to construct covariates. Second, various aspects of SI program implementation,

such as the amount and quality of professional development activities, undoubtedly affected the kinds of teaching practices that were used. Even if two teachers report using reform practices with similar frequency, their approaches to those practices may differ substantially and may reflect specific features of the local reform program. Third, the achievement measures used at each site varied on a number of dimensions, including psychometric quality (e.g., reliability), content, and degree of alignment with the local curriculum.

This last source of differences has implications for future evaluations of SIs and other reforms. Most evaluations rely on locally available student achievement data, in large part because administering additional measures is expensive and often not feasible. Many principals and teachers believe that their students spend far too much time taking the tests that are required locally, and they are therefore reluctant to volunteer for supplementary testing. Locally developed and administered tests may also be preferred because they are presumed to be more closely aligned with local reform efforts than a measure chosen and administered by outside evaluators would be. Although many districts and states are working to develop tests that reflect local standards and curricula, at the sites we studied, most test development lagged far behind reform implementation, leaving local personnel to rely on tests they did not necessarily believe were appropriate.

Our analyses revealed that in most cases, tests that we added specifically for the Mosaic study functioned better than the locally administered tests. Local tests often had lower reliability than our tests. In addition, local tests exhibited unexpected relationships with other measures of student achievement and with student background characteristics that raise questions about the validity of scores. For example, at two sites, there was an extremely high correlation between the percentage of students in a school receiving free or reduced-price lunch and the school's average score on our supplementary multiple-choice and open-ended tests (which is consistent with nearly all prior research on this topic), but the correlation with the local test, measuring similar content, was close to zero. This unexpected result suggests that caution is warranted in interpreting the results of the locally administered tests.

Although the overall differences we observed between achievement on multiple-choice and open-response tests were not significant, the general pattern suggests that format effects should be investigated further. In particular, it raises questions concerning whether the two types of tests measure different constructs. Most advocates of systemic reform believe that traditional multiple-choice tests do not adequately reflect the range of competencies the reforms are expected to develop, and that tests requiring students to construct their answers and to engage in complex problem-solving are more appropriate. Our results do not indicate that this is necessarily the case, but the question deserves further investigation, particularly given the resources that many states and districts are devoting to open-ended testing.

PLANS FOR FUTURE DATA COLLECTION AND ANALYSIS

During the second year of the Mosaic project, data were collected from one of the sites that participated during the first year and five additional sites. The data-collection design was essentially the same as that of the first year and will provide us with additional estimates of the relationships between teaching practices and student outcomes. At the site that was also used in year 1, the same grade level was used, so we will not be able to track students over time. However, results from this site will enable us to explore changes in teaching practices by individual teachers and to examine what happens to student achievement as schools and teachers are involved in the reform for longer periods of time.

We have also planned a longitudinal study using one of the sites that participated during year 2. At this site, we hope to follow students over three years, collecting data on the instructional practices used by their teachers during each year. This will enable us to conduct a multiyear "dose-response" study in which degree of exposure to the practices can be related to student growth in achievement over a longer period of time.

PARTICIPATION AT YEAR 1 SITES

SITE 1

Our sample at site 1 consists of fourth-grade students from 17 schools. Our original sample consisted of 20 schools, but two schools refused to participate, one nominated as highly involved and one nominated as a control. We also excluded a school where no third-grade teachers responded. This school was also nominated as highly involved in the reform. From the remaining 17 schools, we obtained survey responses from 46 of the 60 third-grade teachers we surveyed.

We linked 1,012 fourth-grade students to the responding teachers. For our analyses, we eliminated students who were missing scores on the fourth-grade test, as well as a handful who were missing data on other covariates. This left 804 students for our analyses.

SITE 2

Our site 2 sample contains students from 20 schools—10 schools nominated as highly engaged in the reform and 10 matched control schools. The site's systemic reform was implemented for both mathematics and science, and schools involved in one reform were expected to be involved in the other. Hence, we used the same schools to study the effects of both mathematics and science teaching practices.

We surveyed 115 fifth-grade teachers in these 20 schools. We asked every teacher about teaching practices for both mathematics and

science and obtained responses for 100 mathematics and 99 science teachers.

There were 2,345 fifth graders in these 20 schools.[1] However, we could not accurately link 45 students from one school to their science teacher, because teachers at this school shared teaching responsibility in an informal manner that was not documented. We excluded these students and their teachers from our analyses. In addition, we excluded 444 students who did not complete the fifth-grade state multiple-choice tests. We also excluded 21 students for whom we had incomplete data on background characteristics. Of the remaining 1,835 students, 1,639 completed both of the hands-on science tasks and 1,662 completed the multiple-choice science test.

Our two outcome measures for mathematics achievement were the state's fifth-grade multiple-choice mathematics test and a RAND-administered open-ended mathematics test. Of the 2,345 students in our sample, we eliminated 45 students whose links to teachers could not be verified. We also excluded 444 students who did not complete fifth-grade state testing. An additional 19 students had incomplete data on background characteristics. Of the remaining 1,837 students, 1,651 completed the SAT-9 open-ended tests. We excluded 151 students who completed the state's multiple-choice mathematics test but received a score of zero, because they appear to be outliers in a number of respects (although including them has no substantial impact on results). Therefore, the sample for our analyses of the state's multiple-choice mathematics tests contains 1,681 students.

We imputed values for missing prior-year test-score data. For the mathematics multiple-choice sample, we imputed 210 prior-year mathematics and reading scores. For the open-ended mathematics sample, we imputed 202 prior-year mathematics and reading scores. For the hands-on science sample, we imputed 200 prior-year mathematics and reading scores. For the multiple-choice science sample, we imputed 203 prior-year mathematics and reading scores. We used hierarchical Bayesian models (Schafer, 1997) to impute mul-

[1] Fifth-grade classrooms are any classrooms that contain fifth-grade students. Because several schools in our site 2 sample use multigrade classes, some of our fifth-grade classrooms include fourth- or sixth-grade students as well as fifth graders. However, only fifth-grade students are included in our study.

tiple values for each missing value. The imputation models included all variables used in our regression models and contemporaneous reading and mathematics scores. The models also accounted for the hierarchical structure of the data for students nested within classrooms.

SITE 3

Our site 3 mathematics sample consists of students from 18 schools throughout the school district—10 schools nominated as highly engaged in the reform and eight matched control schools. Our original sample consisted of 20 schools, 11 nominated as highly engaged and nine matched controls. However, we obtained no teacher survey responses from two schools, so data from these schools were excluded from our analyses. Our science sample from site 3 consists of 20 schools from the district—10 nominated as highly engaged and 10 matched controls. Five schools are included in both the mathematics and science samples.

We obtained survey responses from 73 of the 87 mathematics teachers in our sample and 74 of the 85 science teachers. Only students whose teachers responded were included in our study.

There were 1,498 eligible fifth graders in our mathematics sample. We excluded students whose teacher did not complete a survey; students exempted from district testing because of LEP or special education status; and students not in fifth grade, even if they were in mixed grade-level classrooms. We included all students identified as LEP or special education but not exempted from district testing. We believe that if the district uses these students' scores to measure school outcomes, the scores are appropriate for measuring the effects of teaching practices. Of the 1,498 eligible students, 1,366 completed the open-ended mathematics test and 1,451 completed both the general mathematics and the computation subtests of the district's multiple-choice mathematics test. An additional 16 students completed only the general mathematics portion, and six students completed only the computation portion of the test.

We used identical criteria to create the science sample, leaving us with 1,652 eligible fifth graders. Of these 1,652 students, 1,438 were administered the multiple-choice science test and 1,367 were administered a hands-on science test.

We imputed values for missing prior-year test-score data. For the mathematics multiple-choice sample, we imputed 390 prior-year mathematics scores and 412 prior-year reading scores. For the open-ended mathematics sample, we imputed 380 prior-year mathematics scores and 401 prior-year reading scores. For the hands-on science sample, we imputed 334 prior-year mathematics scores and 353 prior-year reading scores. For the multiple-choice science sample, we imputed 346 prior-year mathematics scores and 366 prior-year reading scores. We used hierarchical Bayesian models (Schafer, 1997) to impute multiple values for each missing value. The imputation models included all variables used in our regression models and contemporaneous reading and mathematics scores. The models also accounted for the hierarchical structure of the data for students nested within classrooms.

SITE 4

Our original site 4 sample included 25 schools, 11 nominated as high implementing and 14 as controls. There were 74 eligible teachers teaching 1,566 eligible students in these 25 schools. Four schools in the original sample did not conduct science testing, and no teachers responded from an additional two schools in the sample. Thus, our final sample included 19 schools, 62 eligible teachers, and 1,314 eligible students. From this sample, a total of 49 teachers responded to our survey.

Only 45 of the 49 teachers who completed the survey also conducted science testing, and those teachers make up our final sample. This sample contains 1,012 students, of whom only 954 completed any science tests: 932 completed the multiple-choice test and 909 completed the hands-on tasks.

We imputed values for missing prior-year test-score data. For the multiple-choice sample, we imputed 41 prior-year mathematics and reading scores. For the hands-on science sample, we imputed 38 prior-year mathematics and reading scores. We used hierarchical Bayesian models (Schafer, 1997) to impute multiple values for each missing value. The imputation models included all variables used in our regression models and contemporaneous reading and mathematics scores. The models also accounted for the hierarchical structure of the data for students nested within classrooms.

SITE 5

At site 5, we started with a sample of 20 schools in which we evaluated mathematics instruction (10 nominated as engaged in the reform and 10 nominated as yet to be involved) and a separate sample of 20 schools (10 engaged and 10 yet to be involved) for science instruction. Two schools were in both the mathematics and science samples. Three schools refused to participate in our mathematics study, and one refused to be in our science study.

We obtained survey responses from 48 of the 49 mathematics teachers in our mathematics sample and 33 of the 46 science teachers in our science sample. Only students whose teachers responded were eligible for our study.

Our mathematics sample consisted of 3,199 students in the participating schools, 2,940 of whom completed the open-ended mathematics test and 3,028 of whom completed the state's multiple-choice mathematics test. Some students were not tested because of absence from school or because they are exempted from testing (e.g., some special education students). For our analyses of the open-response test scores, we excluded 48 students who received a score of zero on the open-ended test.[2] Hence, our final analysis samples included 2,937 students with open-ended mathematics test scores and 3,018 students with multiple-choice mathematics scores.

Our science sample included 2,436 students in the participating schools. We excluded two students who received a score of zero on the multiple-choice science test. Of the eligible students, 2,047 completed both of the hands-on science tasks and 2,079 completed the multiple-choice science test.

For the multiple-choice mathematics sample, we imputed free or reduced-price lunch status for 60 students, age for 26 students, and contemporaneous language test scores for 10 students. For the open-response mathematics sample, we imputed free or reduced-price lunch status for 136 students, age for 11 students, and contemporaneous language test scores for 95 students. For the multiple-

[2]Students who scored zero on the open-ended math test were outliers and influential in our results. However, we did not want our conclusions to be sensitive to a handful of unrepresentative students, so we excluded these students from our analysis.

choice science sample, we imputed free or reduced-price lunch status for 53 students, age for one student, and contemporaneous language test scores for 34 students. For the hands-on science sample, we imputed free or reduced-price lunch status for 54 students, age for 17 students, and contemporaneous language test scores for 36 students. For the science multiple-choice sample, we imputed free or reduced-price lunch status for 53 students, age for one student, and contemporaneous language test scores for 34 students. For the hands-on science sample, we imputed free or reduced-price lunch status for 54 students, age for 17 students, and contemporaneous language test scores for 36 students. We used hierarchical Bayesian models (Schafer, 1997) to impute multiple values for each missing value. The imputation models included all variables used in our regression models. The models also accounted for the hierarchical structure of the data for students nested within classrooms.

SITE 6

Our site 6 sample included all 25 middle schools in the district. Twelve schools had been involved in the Urban Systemic Initiative (USI) for one or more years, and 13 had not yet been involved. All 58 seventh-grade mathematics and 57 seventh-grade science teachers identified by the district completed surveys. However, only 57 of the responding mathematics teachers linked to seventh-grade students, and only 52 of the science teachers linked to students. The remaining respondents appeared to be incorrectly identified as seventh-grade teachers. The science sample included 3,812 students, and 3,279 of these had scores on the multiple-choice science test. The mathematics sample included 3,682 students, of whom 3,237 had scores on the multiple-choice mathematics test.

For the science sample, we imputed prior-year science test scores for 445 students. For the mathematics sample, we imputed prior-year mathematics test scores for 407 students. We used hierarchical Bayesian models (Schafer, 1997) to impute multiple values for each missing value. The imputation models included all variables used in our regression models. We also included sixth- and seventh-grade reading scores in the imputation models, and all models included both sixth- and seventh-grade science and mathematics scores. The models also accounted for the hierarchical structure of the data for students nested within classrooms.

Appendix B

ITEMS ON TEACHING-PRACTICES SCALES

The wording of items on the teaching-practices scales is identical across sites. The reform-practices score for mathematics is the sum of scores on the 22 items listed in Table B.1, and the traditional-practices score for mathematics is the sum of scores on the five items listed in Table B.2.

The reform-practices score for science is the sum of scores on the 22 items listed in Table B.3, and the traditional-practices score for science is the sum of scores on the five items listed in Table B.4.

Table B.1

Items on Reform-Practices Scale for Mathematics

About how often do you typically do each of the following in your *mathematics* instruction in this class?

- Arrange seating to facilitate student discussion
- Use open-ended questions
- Require students to explain their reasoning when giving an answer
- Encourage students to communicate mathematically
- Encourage students to explore alternative methods for solutions
- Allow students to work at their own pace
- Read and comment on the reflections students have written in their notebooks or journals

About how often do students in this class typically take part in each of the following activities as part of their *mathematics* instruction?

- Participate in student-led discussions
- Work in cooperative learning groups
- Make formal presentations to the class
- Work on solving a real-world problem
- Share ideas or solve problems with each other in small groups
- Engage in hands-on mathematical activities
- Design or implement their *own* investigations
- Work on extended mathematics investigations (a week or more in duration)
- Participate in fieldwork
- Record, represent, and/or analyze data
- Write a description of a plan, procedure, or problem-solving process
- Write reflections in a notebook or journal
- Work on portfolios
- Take tests requiring open-ended responses (e.g., descriptions, justifications of solutions)
- Engage in performance tasks for assessment purposes

Table B.2

Items on Traditional-Practices Scale for Mathematics

About how often do you typically do each of the following in your *mathematics* instruction in this class?

- Lecture/introduce content through formal presentations

About how often do students in this class typically take part in each of the following activities as part of their *mathematics* instruction?

- Read from a mathematics textbook in class
- Practice computational skills
- Memorize mathematics facts, rules, or formulas
- Take short-answer tests (e.g., multiple-choice, true/false, fill-in-the-blank)

Table B.3

Items on Reform-Practices Scale for Science

About how often do you typically do each of the following in your *science* instruction in this class?

- Arrange seating to facilitate student discussion
- Use open-ended questions
- Require students to supply evidence to support their claims
- Encourage students to explain concepts to one another
- Encourage students to consider alternative explanations
- Allow students to work at their own pace
- Read and comment on the reflections students have written in their notebooks or journals

About how often do students in this class typically take part in each of the following activities as part of their *science* instruction?

- Participate in student-led discussions
- Work in cooperative learning groups
- Make formal presentations to the class
- Work on solving a real-world problem
- Share ideas or solve problems with each other in small groups
- Engage in hands-on science activities
- Design or implement their *own* investigations
- Design objects within constraints (e.g., egg drop, toothpick bridge, aluminum boats)
- Work on extended science investigations or projects (a week or more in duration)
- Participate in fieldwork
- Record, represent, and/or analyze data
- Write reflections in a notebook or journal
- Work on portfolios
- Take tests requiring open-ended responses (e.g., descriptions, justifications of solutions)
- Engage in performance tasks for assessment purposes

Table B.4

Items on Traditional-Practices Scale for Science

About how often do you typically do each of the following in your *science* instruction in this class?

Lecture/introduce content through formal presentations

About how often do students in this class typically take part in each of the following activities as part of their *science* instruction?

Read from a science textbook in class

Answer textbook/worksheet questions

Learn science vocabulary

Take short-answer tests (e.g., multiple-choice, true/false, fill-in-the-blank)

FULL REGRESSION MODELS

FULL REGRESSION MODELS FOR SITE 1

Table C.1

Site 1 Regression Models for Mathematics Tests (Grade 4)

State Open-Response Mathematics Test, N = 804

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.289	3.130	0.002
Free or reduced-price lunches	–0.189	–2.856	0.004
African-American[a]	–0.140	–1.770	0.077
Hispanic	–0.252	–2.345	0.019
Other race	–0.144	–1.267	0.205
Female	–0.023	–0.339	0.735
Reading score	0.487	12.927	0.000
Reform scale[b]	–0.010	–0.261	0.794
Model Including Traditional-Practices Scale			
Intercept	0.294	3.266	0.001
Free or reduced-price lunches	–0.192	–2.957	0.003
African-American[a]	–0.141	–1.806	0.071
Hispanic	–0.252	–2.378	0.017
Other race	–0.144	–1.268	0.205
Female	–0.022	–0.332	0.740
Reading score	0.486	12.832	0.000
Traditional scale[b]	–0.006	–0.156	0.876

[a]Reference group for ethnicity is white, non-Hispanic.

[b]Standardized to mean = 0, standard deviation = 1.

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

Table C.1 (continued)

State Grid-In Mathematics Test, N = 804

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.328	4.863	0.000
Free or reduced-price lunches	–0.159	–1.914	0.056
African-American[a]	–0.253	–3.730	0.000
Hispanic	–0.152	–1.423	0.155
Other race	–0.090	–0.834	0.404
Female	–0.071	–1.098	0.272
Reading score	0.516	16.964	0.000
Reform scale[b]	0.050	1.941	0.052
Model Including Traditional-Practices Scale			
Intercept	0.324	4.749	0.000
Free or reduced-price lunches	–0.157	–1.872	0.061
African-American[a]	–0.252	–3.726	0.000
Hispanic	–0.166	–1.521	0.128
Other race	–0.093	–0.870	0.385
Female	–0.072	–1.113	0.266
Reading score	0.514	16.832	0.000
Traditional scale[b]	–0.040	–1.792	0.073

State Multiple-Choice Mathematics Test, N = 804

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.462	8.821	0.000
Free or reduced-price lunches	–0.206	–3.387	0.001
African-American[a]	–0.264	–3.709	0.000
Hispanic	–0.160	–1.986	0.047
Other race	–0.218	–1.852	0.064
Female	–0.224	–4.099	0.000
Reading score	0.631	19.294	0.000
Reform scale[b]	0.031	0.870	0.385
Model Including Traditional-Practices Scale			
Intercept	0.463	9.206	0.000
Free or reduced-price lunches	–0.207	–3.303	0.001
African-American[a]	–0.265	–3.864	0.000
Hispanic	–0.172	–2.053	0.040
Other race	–0.221	–1.872	0.061
Female	–0.224	–4.129	0.000
Reading score	0.628	19.014	0.000
Traditional scale[b]	–0.037	–1.223	0.221

[a]Reference group for ethnicity is white, non-Hispanic.

[b]Standardized to mean = 0, standard deviation = 1.

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

FULL REGRESSION MODELS FOR SITE 2

Table C.2

Site 2 Regression Models for Science and Mathematics Tests (Grade 5)

RAND Hands-On Science Tasks, N = 1,639

Predictor Variable	Coefficient	T-statistic[d]	p-value
Model Including Reform-Practices Scale			
Intercept	0.183	2.361	0.018
Free or reduced-price lunches	–0.195	–3.818	0.000
LEP	0.025	0.380	0.704
Special education	–0.422	–4.975	0.000
Gifted	0.373	6.108	0.000
Minority[a]	–0.081	–1.143	0.253
Female	0.099	2.217	0.027
Grade 4 state MC math[b]	0.279	9.231	0.000
Grade 4 state MC reading[b]	0.291	10.207	0.000
Reform scale[c]	0.094	3.511	0.000
Model Including Traditional-Practices Scale			
Intercept	0.195	2.561	0.010
Free or reduced-price lunches	–0.198	–3.712	0.000
LEP	0.009	0.125	0.900
Special education	–0.409	–4.925	0.000
Gifted	0.376	5.984	0.000
Minority[a]	–0.098	–1.335	0.182
Female	0.101	2.247	0.025
Grade 4 state MC math[b]	0.276	9.165	0.000
Grade 4 state MC reading[b]	0.291	10.345	0.000
Traditional scale[c]	0.035	1.192	0.233

[a]Includes all students not identified as white, non-Hispanic.
[b]NCE standardized to mean = 0, standard deviation = 1.
[c]Standardized to mean = 0, standard deviation = 1.
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

Table C.2 (continued)

Stanford 9 Multiple-Choice Science Test, N = 1,662

Predictor Variable	Coefficient	T-statistic[d]	p-value
Model Including Reform-Practices Scale			
Intercept	0.392	5.499	0.000
Free or reduced-price lunches	–0.107	–1.914	0.056
LEP	–0.029	–0.467	0.641
Special education	–0.550	–4.752	0.000
Gifted	0.309	5.480	0.000
Minority[a]	–0.208	–2.813	0.005
Female	–0.188	–4.018	0.000
Grade 4 state MC math[b]	0.161	4.936	0.000
Grade 4 state MC reading[b]	0.325	11.103	0.000
Reform scale[c]	0.054	2.043	0.041
Model Including Traditional-Practices Scale			
Intercept	0.394	5.411	0.000
Free or reduced-price lunches	–0.106	–1.819	0.069
LEP	–0.044	–0.694	0.488
Special education	–0.550	–4.753	0.000
Gifted	0.306	5.526	0.000
Minority[a]	–0.214	–2.793	0.005
Female	–0.187	–3.976	0.000
Grade 4 state MC math[b]	0.159	4.994	0.000
Grade 4 state MC reading[b]	0.326	11.398	0.000
Traditional scale[c]	0.041	1.522	0.128

Stanford 9 Open-Response Mathematics Test, N = 1,651

Predictor Variable	Coefficient	T-statistic[d]	p-value
Model Including Reform-Practices Scale			
Intercept	0.180	1.705	0.088
Free or reduced-price lunches	–0.078	–1.587	0.113
LEP	–0.036	–0.459	0.646
Special education	–0.385	–3.996	0.000
Gifted	0.426	6.960	0.000
Minority[a]	–0.074	–0.810	0.418
Female	–0.103	–2.334	0.020
Grade 4 state MC math[b]	0.302	7.540	0.000
Grade 4 state MC reading[b]	0.146	3.995	0.000
Reform scale[c]	0.052	1.244	0.214
Model Including Traditional-Practices Scale			
Intercept	0.181	1.779	0.075
Free or reduced-price lunches	–0.081	–1.702	0.089
LEP	–0.035	–0.469	0.639
Special education	–0.391	–4.158	0.000
Gifted	0.425	7.149	0.000
Minority[a]	–0.074	–0.811	0.417
Female	–0.103	–2.340	0.019
Grade 4 state MC math[b]	0.304	7.667	0.000
Grade 4 state MC reading[b]	0.141	3.828	0.000
Traditional scale[c]	0.000	–0.012	0.991

[a]Includes all students not identified as white, non-Hispanic.
[b]NCE standardized to mean = 0, standard deviation = 1.
[c]Standardized to mean = 0, standard deviation = 1.
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

Table C.2 (continued)

State Multiple-Choice Mathematics Test, N = 1,686

Predictor Variable	Coefficient	T-statistic[d]	p-value
Model Including Reform-Practices Scale			
Intercept	0.028	0.303	0.762
Free or reduced-price lunches	0.008	0.147	0.883
LEP	0.075	0.956	0.339
Special education	–0.527	–4.199	0.000
Gifted	0.170	3.156	0.002
Minority[a]	–0.030	–0.478	0.633
Female	0.026	0.652	0.514
Grade 4 state MC math[b]	0.543	13.136	0.000
Grade 4 state MC reading[b]	0.116	3.679	0.000
Reform scale[c]	0.023	0.618	0.536
Model Including Traditional-Practices Scale			
Intercept	0.037	0.432	0.666
Free or reduced-price lunches	–0.001	–0.009	0.993
LEP	0.045	0.603	0.546
Special education	–0.514	–4.197	0.000
Gifted	0.157	2.867	0.004
Minority[a]	–0.021	–0.354	0.723
Female	0.028	0.700	0.484
Grade 4 state MC math[b]	0.539	13.157	0.000
Grade 4 state MC reading[b]	0.119	3.701	0.000
Traditional scale[c]	0.088	2.195	0.028

[a]Includes all students not identified as white, non-Hispanic.

[b]NCE standardized to mean = 0, standard deviation = 1.

[c]Standardized to mean = 0, standard deviation = 1.

[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

FULL REGRESSION MODELS FOR SITE 3

Table C.3

Site 3 Regression Models for Science and Mathematics Tests (Grade 5)

CSIAC Hands-On Science Test, N = 1,367

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.139	1.797	0.072
Free or reduced-price lunches	–0.106	–1.305	0.192
English proficiency 1	0.062	0.289	0.774
English proficiency 2	0.041	0.492	0.623
Special education	–0.146	–1.435	0.151
African-American[a]	–0.074	–0.865	0.387
Hispanic	–0.155	–2.818	0.005
Asian	0.028	0.237	0.813
Female	0.205	4.376	0.000
Grade 4 math	0.172	4.232	0.000
Grade 4 reading	0.352	7.731	0.000
Missing grade 4 scores	–0.160	–2.104	0.035
Reform scale[b]	0.079	1.281	0.200
Model Including Traditional-Practices Scale			
Intercept	0.140	1.737	0.082
Free or reduced-price lunches	–0.113	–1.319	0.187
English proficiency 1	0.069	0.320	0.751
English proficiency 2	0.043	0.519	0.604
Special education	–0.136	–1.350	0.177
African-American[a]	–0.065	–0.736	0.462
Hispanic	–0.145	–2.650	0.008
Asian	0.028	0.242	0.808
Female	0.209	4.394	0.000
Grade 4 math	0.175	4.273	0.000
Grade 4 reading	0.358	7.684	0.000
Missing grade 4 scores	–0.153	–2.014	0.044
Traditional scale[b]	0.075	1.328	0.184

[a]Reference group for ethnicity is white, non-Hispanic.

[b]Standardized to mean = 0, standard deviation = 1.

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

Table C.3 (continued)

CSIAC Multiple-Choice Science Test, N = 1,438

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.473	8.291	0.000
Free or reduced-price lunches	–0.190	–3.388	0.001
English proficiency 1	–0.199	–0.855	0.405
English proficiency 2	–0.057	–0.804	0.421
Special education	0.022	0.336	0.737
African-American[a]	–0.248	–3.336	0.001
Hispanic	–0.193	–3.549	0.000
Asian	–0.239	–2.756	0.006
Female	–0.078	–1.714	0.087
Grade 4 math	0.135	3.813	0.000
Grade 4 reading	0.521	16.222	0.000
Missing grade 4 scores	–0.128	–2.389	0.017
Reform scale[b]	0.049	1.951	0.051
Model Including Traditional-Practices Scale			
Intercept	0.480	8.247	0.000
Free or reduced-price lunches	–0.202	–3.568	0.000
English proficiency 1	–0.192	–0.821	0.423
English proficiency 2	–0.057	–0.806	0.420
Special education	0.024	0.377	0.706
African-American[a]	–0.242	–3.286	0.001
Hispanic	–0.193	–3.470	0.001
Asian	–0.241	–2.746	0.006
Female	–0.078	–1.709	0.088
Grade 4 math	0.136	3.806	0.000
Grade 4 reading	0.526	16.461	0.000
Missing grade 4 scores	–0.124	–2.321	0.021
Traditional scale[b]	0.007	0.280	0.779

[a]Reference group for ethnicity is white, non-Hispanic.

[b]Standardized to mean = 0, standard deviation = 1.

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

Table C.3 (continued)

Stanford 9 Open-Response Mathematics Test, N = 1,366

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.356	3.454	0.001
Free or reduced-price lunches	–0.106	–1.171	0.242
English proficiency 1	0.005	0.037	0.970
English proficiency 2	0.134	2.121	0.034
Special education	–0.204	–2.510	0.012
African-American[a]	–0.331	–4.526	0.000
Hispanic	–0.183	–2.915	0.004
Asian	0.026	0.262	0.794
Female	–0.168	–3.407	0.001
Grade 4 math	0.277	5.531	0.000
Grade 4 reading	0.275	6.424	0.000
Missing grade 4 scores	–0.137	–2.296	0.022
Reform scale[b]	0.092	2.055	0.040
Model Including Traditional-Practices Scale			
Intercept	0.374	3.606	0.000
Free or reduced-price lunches	–0.125	–1.386	0.166
English proficiency 1	0.010	0.070	0.944
English proficiency 2	0.135	2.104	0.035
Special education	–0.219	–2.683	0.008
African-American[a]	–0.336	–4.328	0.000
Hispanic	–0.192	–2.914	0.004
Asian	0.043	0.412	0.680
Female	–0.175	–3.560	0.000
Grade 4 math	0.281	5.183	0.000
Grade 4 reading	0.287	6.619	0.000
Missing grade 4 scores	–0.157	–2.489	0.013
Traditional scale[b]	–0.091	–1.760	0.078

[a]Reference group for ethnicity is white, non-Hispanic.

[b]Standardized to mean = 0, standard deviation = 1.

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

Table C.3 (continued)

CTBS Multiple-Choice Mathematics Test, N = 1,451

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.360	4.491	0.000
Free or reduced-price lunches	–0.176	–2.436	0.015
English proficiency 1	–0.036	–0.240	0.812
English proficiency 2	0.102	1.850	0.064
Special education	–0.228	–3.149	0.002
African-American[a]	–0.080	–1.323	0.186
Hispanic	–0.120	–2.318	0.021
Asian	0.044	0.532	0.595
Female	–0.089	–2.059	0.040
Grade 4 math	0.365	9.004	0.000
Grade 4 reading	0.380	13.073	0.000
Missing grade 4 scores	–0.132	–2.377	0.018
Reform scale[b]	0.044	1.047	0.295
Model Including Traditional-Practices Scale			
Intercept	0.366	4.798	0.000
Free or reduced-price lunches	–0.182	–2.464	0.014
English proficiency 1	–0.037	–0.252	0.803
English proficiency 2	0.102	1.868	0.062
Special education	–0.233	–3.195	0.002
African-American[a]	–0.080	–1.288	0.198
Hispanic	–0.121	–2.245	0.025
Asian	0.055	0.640	0.522
Female	–0.093	–2.219	0.027
Grade 4 math	0.368	8.828	0.000
Grade 4 reading	0.386	13.142	0.000
Missing grade 4 scores	–0.143	–2.487	0.014
Traditional scale[b]	–0.050	–1.460	0.144

[a]Reference group for ethnicity is white, non-Hispanic.

[b]Standardized to mean = 0, standard deviation = 1.

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

FULL REGRESSION MODELS FOR SITE 4

Table C.4

Site 4 Regression Models for Science Tests (Grade 5)

CSIAC Hands-On Science Test, N = 909

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.196	2.150	0.032
Free or reduced-price lunches	–0.087	–0.819	0.413
English proficiency 1	–0.401	–2.659	0.008
English proficiency 2	0.438	2.496	0.013
Gifted	0.205	3.044	0.002
Special education	–0.001	–0.011	0.991
African-American[a]	–0.872	–6.630	0.000
Chinese	–0.035	–0.400	0.689
Filipino	–0.316	–1.829	0.067
Hispanic	–0.099	–0.794	0.427
Japanese	–0.084	–0.443	0.658
Korean	0.169	1.085	0.278
Other race/ethnicity	–0.187	–1.529	0.126
Female	0.066	1.016	0.309
Grade 4 math	0.149	1.709	0.088
Grade 4 reading	0.183	1.988	0.047
Reform scale[b]	0.046	0.698	0.485
Model Including Traditional-Practices Scale			
Intercept	0.195	2.102	0.036
Free or reduced-price lunches	–0.080	–0.761	0.446
English proficiency 1	–0.400	–2.610	0.009
English proficiency 2	0.452	2.595	0.009
Gifted	0.203	2.979	0.003
Special education	–0.009	–0.070	0.944
African-American[a]	–0.857	–6.296	0.000
Chinese	–0.028	–0.309	0.757
Filipino	–0.298	–1.694	0.090
Hispanic	–0.100	–0.810	0.418
Japanese	–0.049	–0.260	0.795
Korean	0.182	1.155	0.248
Other race/ethnicity	–0.172	–1.357	0.175
Female	0.064	1.013	0.311
Grade 4 math	0.149	1.728	0.084
Grade 4 reading	0.182	2.008	0.045
Traditional scale[b]	–0.008	–0.138	0.891

[a]Reference group for ethnicity is white, non-Hispanic.
[b]Standardized to mean = 0, standard deviation = 1.
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

Table C.4 (continued)

CSIAC Multiple-Choice Science Test, N = 932

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.219	2.793	0.005
Free or reduced-price lunches	–0.155	–2.519	0.012
English proficiency 1	–0.483	–5.645	0.000
English proficiency 2	0.216	1.247	0.212
Gifted	0.347	5.633	0.000
Special education	0.214	1.898	0.058
African-American[a]	–0.661	–7.437	0.000
Chinese	0.002	0.024	0.981
Filipino	–0.244	–2.193	0.028
Hispanic	–0.037	–0.368	0.713
Japanese	0.010	0.049	0.961
Korean	0.317	3.232	0.001
Other race/ethnicity	–0.110	–1.607	0.108
Female	–0.232	–5.162	0.000
Grade 4 math	0.114	2.050	0.041
Grade 4 reading	0.407	8.041	0.000
Reform scale[b]	–0.003	–0.094	0.925
Model Including Traditional-Practices Scale			
Intercept	0.144	1.817	0.069
Free or reduced-price lunches	–0.156	–2.771	0.006
English proficiency 1	–0.479	–5.387	0.000
English proficiency 2	0.241	1.349	0.177
Gifted	0.338	5.343	0.000
Special education	0.217	1.892	0.059
African-American[a]	–0.682	–7.143	0.000
Chinese	0.006	0.097	0.923
Filipino	–0.244	–2.090	0.037
Hispanic	–0.056	–0.532	0.595
Japanese	0.049	0.193	0.847
Korean	0.312	3.078	0.002
Other race/ethnicity	–0.108	–1.540	0.123
Female	–0.218	–4.610	0.000
Grade 4 math	0.120	2.202	0.028
Grade 4 reading	0.397	7.962	0.000
Traditional scale[b]: linear	–0.052	–1.171	0.242
Traditional scale[b]: quadratic	0.077	1.985	0.047

[a]Reference group for ethnicity is white, non-Hispanic.

[b]Standardized to mean = 0, standard deviation = 1.

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

FULL REGRESSION MODELS FOR SITE 5

Table C.5

Site 5 Regression Models for Science and Mathematics Tests (Grade 7)

RAND Hands-On Science Tasks, N = 2,047

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	1.338	2.721	0.007
Free or reduced-price lunches	–0.180	3.505	0.000
Age	–0.088	2.285	0.022
Minority[a]	–0.549	5.666	0.000
Female	0.038	0.995	0.320
Grade 7 language	0.392	24.769	0.000
Special education	–0.040	0.309	0.757
Reform scale[b]	0.001	0.015	0.988
Model Including Traditional-Practices Scale			
Intercept	1.321	3.264	0.001
Free or reduced-price lunches	–0.122	2.485	0.013
Age	–0.092	2.877	0.004
Minority[a]	–0.529	7.820	0.000
Female	0.024	0.731	0.464
Grade 7 language	0.390	24.221	0.000
Special education	–0.038	0.333	0.739
Traditional scale[b]	–0.166	3.966	0.000

Stanford 9 Multiple-Choice Science Test, N = 2,079

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	0.733	1.285	0.199
Free or reduced-price lunches	–0.135	2.987	0.003
Age	–0.032	0.737	0.461
Minority[a]	–0.518	4.088	0.000
Female	–0.176	3.358	0.001
Grade 7 language	0.526	24.101	0.000
Special education	–0.240	2.144	0.032
Reform scale[b]	0.030	0.377	0.707
Model Including Traditional-Practices Scale			
Intercept	0.688	1.154	0.249
Free or reduced-price lunches	–0.114	2.483	0.013
Age	–0.031	0.674	0.501
Minority[a]	–0.510	4.246	0.000
Female	–0.179	3.393	0.001
Grade 7 language	0.527	22.569	0.000
Special education	–0.216	1.896	0.058
Traditional scale[b]	–0.080	1.622	0.105

[a]Includes all students not identified as white, non-Hispanic.

[b]Standardized to mean = 0, standard deviation = 1.

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

Table C.5 (continued)

Stanford 9 Open-Response Mathematics Tests, N=2,937

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	1.902	5.984	0.000
Free or reduced-price lunches	–0.056	1.712	0.087
Age	–0.130	5.233	0.000
Minority[a]	–0.365	9.719	0.000
Female	–0.128	4.610	0.000
Grade 7 language	0.361	15.145	0.000
Special education	–0.004	0.030	0.976
Reform scale[b]	0.080	2.691	0.007
Model Including Traditional-Practices Scale			
Intercept	1.977	5.868	0.000
Free or reduced-price lunches	–0.064	2.010	0.044
Age	–0.136	5.075	0.000
Minority[a]	–0.377	10.226	0.000
Female	–0.130	4.655	0.000
Grade 7 language	0.364	15.216	0.000
Special education	–0.028	0.190	0.849
Traditional scale[b]	–0.028	0.943	0.346

State Multiple-Choice Mathematics Tests, N=3,018

Predictor Variable	Coefficient	T-statistic[c]	p-value
Model Including Reform-Practices Scale			
Intercept	1.230	3.214	0.001
Free or reduced-price lunches	–0.047	1.654	0.098
Age	–0.084	3.030	0.002
Minority[a]	–0.106	3.036	0.002
Female	–0.165	6.975	0.000
Grade 7 language	0.674	28.019	0.000
Special education	0.176	2.068	0.039
Reform scale[b]	–0.052	0.897	0.369
Model Including Traditional-Practices Scale			
Intercept	1.173	3.085	0.002
Free or reduced-price lunches	–0.045	1.634	0.102
Age	–0.079	2.913	0.004
Minority[a]	–0.102	2.833	0.005
Female	–0.164	7.086	0.000
Grade 7 language	0.673	28.193	0.000
Special education	0.193	2.205	0.027
Traditional scale[b]	0.037	0.660	0.510

[a]Includes all students not identified as white, non-Hispanic.
[b]Standardized to mean = 0, standard deviation = 1.
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

FULL REGRESSION MODELS FOR SITE 6

Table C.6

Site 6 Regression Models for Science and Mathematics Tests (Grade 7)

MAT-7 Multiple-Choice Science Test, N=3,279

Predictor Variable	Coefficient	T-statistic[d]	p-value
Model Including Reform-Practices Scale			
Intercept	0.146	3.472	0.001
Free or reduced-price lunches	–0.188	–5.729	0.000
Minority[a]	–0.062	–1.948	0.051
Female	–0.053	–1.758	0.079
Grade 6 science[b]	0.620	26.376	0.000
Reform scale[c]	0.051	2.750	0.006
Model Including Traditional-Practices Scale			
Intercept	0.149	3.351	0.001
Free or reduced-price lunches	–0.198	–6.071	0.000
Minority[a]	–0.051	–1.620	0.105
Female	–0.053	–1.775	0.076
Grade 6 science[b]	0.624	26.640	0.000
Traditional scale[c]	–0.004	–0.149	0.881

MAT-7 Multiple-Choice Mathematics Test, N=3,237

Predictor Variable	Coefficient	T-statistic[d]	p-value
Model Including Reform-Practices Scale			
Intercept	0.097	1.867	0.062
Free or reduced-price lunches	–0.076	–1.917	0.055
Minority[a]	–0.048	–1.185	0.236
Female	–0.031	–1.308	0.191
Grade 6 math[b]	0.755	24.564	0.000
Missing grade 6 math	–0.148	–3.598	0.000
Reform scale[c]	0.046	1.589	0.112
Model Including Traditional-Practices Scale			
Intercept	0.105	1.963	0.050
Free or reduced-price lunches	–0.071	–1.695	0.090
Minority[a]	–0.058	–1.389	0.165
Female	–0.033	–1.393	0.164
Grade 6 math[b]	0.752	23.682	0.000
Missing grade 6 math	–0.154	–3.724	0.000
Traditional scale[c]	–0.004	–0.096	0.923

[a]Includes all students not identified as white, non-Hispanic.
[b]Scale score standardized to mean = 0, standard deviation = 1.
[c]Standardized to mean = 0, standard deviation = 1.
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997).

Appendix D

DETAILS OF POOLED ANALYSIS OF REGRESSION COEFFICIENTS

We used the following model to obtain our pooled estimates and to determine whether sites were comparable and pooled analyses appropriate. For a given subject, test type, and scale, we assumed that our sample is a random sample of all possible sites and that

$$b_i = \beta + \eta_i + \varepsilon_i ,$$

where b_i denotes the estimated coefficient from the ith site, β denotes the average coefficient across all sites, η_i denotes the deviation of site i from the average, and ε_i denotes sampling error in our estimate b_i as an estimate of $\beta + \eta_i$. The deviations, η_i, are assumed to be normally distributed with mean zero and variance τ^2. The errors, ε_i, are assumed to be normally distributed with mean zero and variance σ_i^2. Error variability differs from site to site depending on the distribution of teacher responses and other covariates, and depending on the residual variance from the regression model fit for each site.

We used the square of the standard error estimates from the individual site analyses as our estimates of the σ_i^2 parameters. Treating these estimates as fixed, we then estimated τ^2 using restricted maximum likelihood (Searle, Cassella, and McCullogh, 1992). This method also provides a confidence interval for our estimate of τ^2. We then estimated the average coefficient, β, as a weighted average of the b_is, where the weight for the ith site is

$$w_i = \frac{1}{\hat{\tau}^2 + \hat{\sigma}_i^2} \Big/ \sum \frac{1}{\hat{\tau}^2 + \hat{\sigma}_j^2}$$

The weighted averages, their standard errors, and lower and upper 95 percent confidence bounds are provided in Table D.1. Estimates and confidence intervals of τ^2 are also given for each set of pooled analyses.

Table D.1

Results from Pooled Analyses of Relationships Between Practices and Achievement

Subject	Test Format	Scale	Weighted Average Coefficient	Standard Error	CI for Coefficient	τ^2	CI for τ^2
Math	OR	Reform	0.053	0.023	(0.008, 0.098)	0.001	(0.000, 0.017)
Math	MC	Reform	0.030	0.017	(–0.003, 0.063)	0.000	(0.000, 0.004)
Science	OR	Reform	0.079	0.022	(0.036, 0.122)	0.000	(0.000, 0.015)
Science	MC	Reform	0.045	0.012	(0.022, 0.069)	0.000	(0.000, 0.003)
Math	OR	Trad.	–0.025	0.019	(–0.061, 0.012)	0.000	(0.000, 0.009)
Math	MC	Trad.	0.001	0.026	(–0.049, 0.052)	0.002	(0.000, 0.018)
Science	OR	Trad.	–0.018	0.054	(–0.123, 0.088)	0.009	(0.001, 0.098)
Science	MC	Trad.	0.006	0.015	(–0.023, 0.034)	0.000	(0.000, 0.013)

Appendix E

RESULTS FROM ANALYSIS OF FORMAT DIFFERENCES

Table E.1

Standardized Coefficients for Models Predicting Differences Between Formats

Site	Subject	Scale	Beta	T-statistic	p-value
1	Math	Reform	–0.041	–1.364	0.179
1	Math	Traditional	0.032	0.906	0.370
2	Math	Reform	0.032	0.664	0.507
2	Math	Traditional	–0.098	–1.862	0.063
2	Science	Reform	0.036	1.291	0.197
2	Science	Traditional	–0.005	–0.174	0.862
3	Math	Reform	0.051	1.266	0.205
3	Math	Traditional	–0.028	–0.599	0.549
3	Science	Reform	0.045	0.646	0.518
3	Science	Traditional	0.071	1.107	0.268
4	Science	Reform	0.048	0.805	0.421
4	Science	Traditional	(a)	(a)	(a)
5	Math	Reform	0.132	2.683	0.010
5	Math	Traditional	–0.060	–1.008	0.319
5	Science	Reform	–0.030	–0.457	0.651
5	Science	Traditional	–0.102	–2.385	0.023

[a]No estimate is available; model for scores and the traditional scale is not linear.

Table E.2

Results from Pooled Analyses of Differences Between Formats

Subject	Scale	Weighted Average	Standard Error	CI Lower Bound	CI Upper Bound	τ^2	CI for τ^2
Math	Reform	0.037	0.037	–0.035	0.109	0.004	(0.000, 0.044)
Science	Reform	0.031	0.022	–0.013	0.075	0.000	(0.000, 0.011)
Math	Trad.	–0.028	0.031	–0.088	0.032	0.002	(0.000, 0.028)
Science	Trad.	–0.019	0.046	–0.109	0.070	0.004	(0.000, 0.119)

Appendix F

SENSITIVITY ANALYSES: USE OF CONTEMPORANEOUS TEST SCORES

The results presented in this report use contemporaneous reading and language scores as covariates in the models for student outcomes at sites 2 and 5 because prior-year test scores were unavailable for these sites. Both prior-year scores and contemporaneous scores serve as measures of student achievement. However, unlike prior-year scores, contemporaneous test scores are not necessarily independent of the teaching practices measured by the survey. If reform teaching in mathematics or science involves activities that promote the use of verbal skills, it is conceivable that students' reading or language scores will be higher when their teachers use greater amounts of reform teaching practices. Including contemporaneous scores in the model might absorb some of the effect of reform (or traditional) practices and could lead to under- or overestimation of the relationship between teaching practices and scores in science or mathematics.

On the other hand, models that exclude contemporaneous scores are probably liberal in the sense that the spurious correlation between student background characteristics that are not included in the model and teacher practices is attributed to the effect of teaching practices. Lacking good independent measures of students' achievement, we chose to present the possibly conservative models that include contemporaneous scores. However, when we use models without contemporaneous scores for sites 1 and 5, the pooled results are similar to the estimates presented in the text, although the estimates without contemporaneous scores tend to be larger (in absolute value). Without contemporaneous scores for sites 1 and 5,

we still estimate a small positive relationship between the reform-practices scale and the mathematics and science multiple-choice and open-response test scores. These estimates are statistically significant except those for multiple-choice mathematics scores.

The results for the traditional scale are similar. For both science multiple-choice and science open-response scores, the pooled results for models without contemporaneous scores are of the same sign and same magnitude as the results from models that include contemporaneous scores. For mathematics, the pooled results for the traditional scale are somewhat different when we exclude the contemporaneous scores. Without contemporaneous scores, the pooled estimate for open-response scores is –0.045 and is statistically significant. With contemporaneous scores, the pooled estimate is –0.025 and is not statistically significant. For both site 1 and site 5, controlling for contemporaneous scores attenuates the negative relationship between the traditional scale and the open-response mathematics scores, although the attenuation is greater for site 1. Similarly, controlling for contemporaneous scores attenuates the negative relationship between the traditional scale and the multiple-choice mathematics scores for sites 1 and 5. The result is that the pooled analysis with estimates from models that include contemporaneous scores yields a very small positive relationship, while the pooled analysis with estimates from models that exclude contemporaneous scores yields a small negative estimate. Neither estimate is statistically significant, and we might conclude that the difference primarily reflects the very weak relationship between the traditional scale and the multiple-choice mathematics scores.

We are uncertain about how the two sets of estimates would compare with estimates that use prior-year scores. Using scores from sites 2, 3, 4, and 6, we found that models with contemporaneous scores often result in attenuated relationships between scores and teaching-practices scales (16 out of 23 models) when compared with models that include prior-year scores. We also found that for these sites, models that include neither prior-year nor contemporaneous scores tend to exaggerate the relationship between the reform scale and the test scores when compared with models that include prior-year scores. However, models that include neither prior-year nor contemporaneous scores tend to attenuate the relationship between the traditional scale and the test scores when compared with models

that include prior-year scores. The difference between traditional and reform scales is due in part to the fact that at some sites, students of teachers who report greater use of traditional practices are somewhat more likely to have lower prior-year scores. Similarly, at some sites, students whose teachers report greater use of reform practices tend to have higher prior-year test scores, although this is less common and the relationship is weaker than the relationship between traditional practices and lower prior-year scores.

Figure F.1 summarizes the results of using prior-year, contemporaneous, or no test scores in our models for mathematics and science for sites 2, 3, 4, and 6. The x-axis is the absolute value of the estimated coefficients from the models that include prior-year test scores. These scores are also plotted as the solid line. The empty squares denote the absolute values for the coefficients for the reform scale in models that do not control for student achievement. The solid squares denote the absolute values for the coefficients for the reform scale in models that use contemporaneous test scores to control for achievement. The empty and solid circles denote the

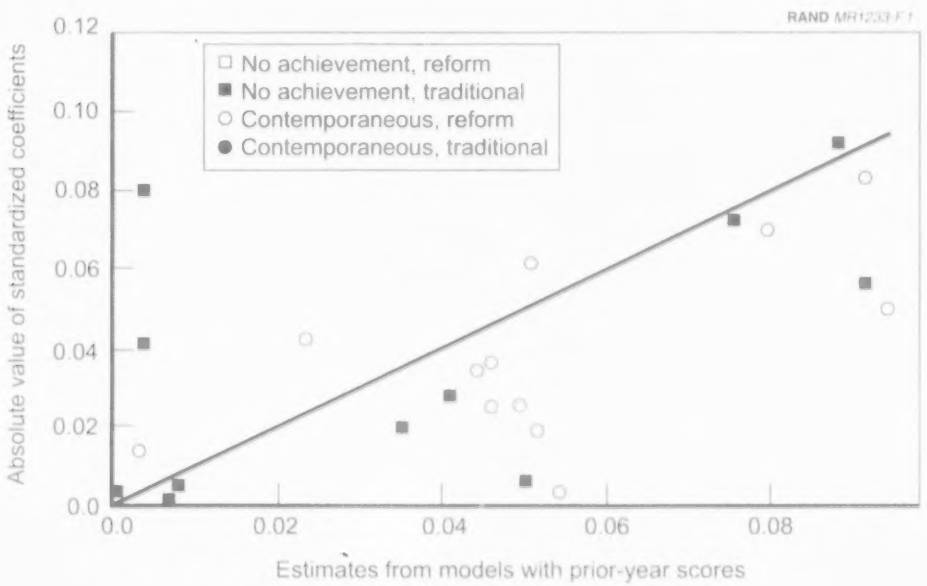


Figure F.1—Estimated Coefficients for Mathematics and Science Tests, Sites 2, 3, 4, and 6: Models Include Prior-Year, Contemporaneous, or No Test Scores

analogous estimates for the traditional scale. Points above the solid line indicate that the estimate was exaggerated compared with the estimate from a model that included prior-year scores. Points below the line indicate that the estimate was attenuated toward zero compared with the estimate from a model that contained the prior-year scores. For example, nine of the 12 empty circles are below the line, indicating that for nine of the 12 models that include the reform scale, including the contemporaneous tests scores attenuates the estimated coefficient for reform compared with including prior-year scores.

SENSITIVITY ANALYSES: COMBINING REFORM AND TRADITIONAL SCALES IN A SINGLE MODEL

Our analyses involved fitting separate models for the reform and traditional scales to estimate the full "effect" of either the reform or the traditional scale. The coefficient for the reform scale estimates the change in the average test score associated with a one-standard-deviation unit increase in the scale, adjusted for differences in student backgrounds. No adjustment is made to account for differences in the use of traditional practices when traditional and reform practices are correlated. In this case, any "effect" due to changes in the traditional scale is attributed to changes in the reform scale. Similarly, the estimated coefficient for the traditional scale is not adjusted for differences in the reform scale. We feel that estimation of the full effect is most interesting because reform practices might encompass both using teaching techniques that are advocated by the reform *and* using fewer traditional practices, or the reform might encompass simply using more of the techniques advocated by the reform. We wanted our estimates to reflect either approach to reform.

Alternatively, we could have included both the reform and traditional scales in our model and estimated the relationship between the reform (traditional) scale and scores, conditional on the level of use of traditional (reform) practices. As discussed above, for most sites the reform and traditional scales are weakly correlated, so including both scales in the same models yields estimates that are very similar to the estimates from fitting separate models. Figure G.1 compares the two sets of estimated coefficients. The standardized coefficients from the separate models are plotted on the x-axis and the solid line. The standardized coefficients from models that include both the tradi-

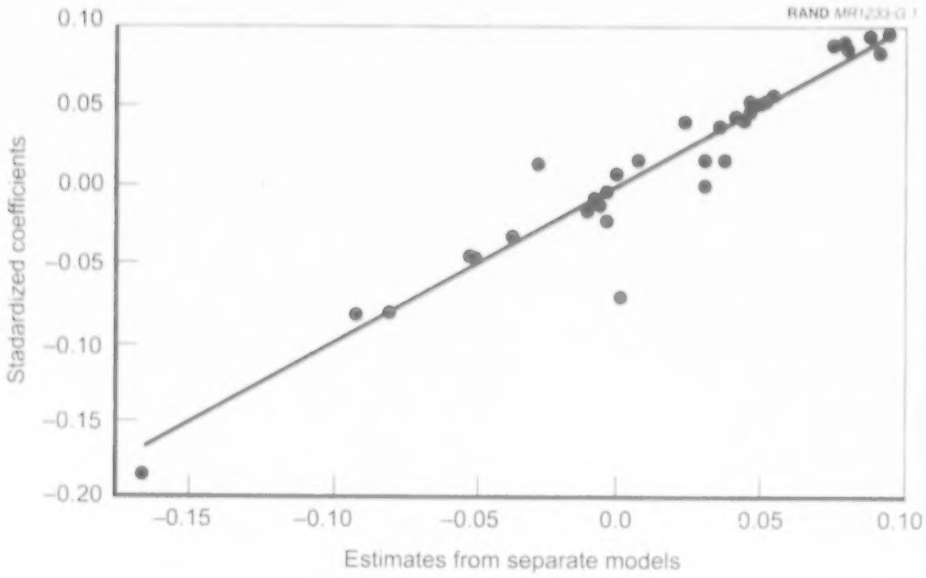


Figure G.1—Estimated Coefficients for Mathematics and Science Tests

tional and reform scales are plotted as the solid circles. The circles closely follow the line, indicating that the estimates from the separate models are very similar to estimates from models that include both the reform scale and the traditional scale. The correlation is 0.958. Only one point deviates from the line: (0.001, -0.070), the estimates for the reform scale and the open-response science test scores for site 5. For this site, the weak positive relationship that we estimate in the model without the traditional scale changes to a moderate negative effect after we control for the traditional scale. The relationship between the traditional scale and the open-response science scores is large and negative for site 5. In addition, teachers who score higher on the reform scale tend to score lower on the traditional scale. Hence, in the model that includes both the traditional and the reform scales, the estimate for the reform scale turns from slightly positive to negative. However, overall, our analyses are not greatly affected by fitting separate models or fitting models that include both scales, although the pooled estimates for the relationship between open-response science scores and the reform scale are somewhat smaller and no longer statistically significant. Figures G.2 through G.9 and Table G.1 provide a summary and pooled results from models that include both scales.

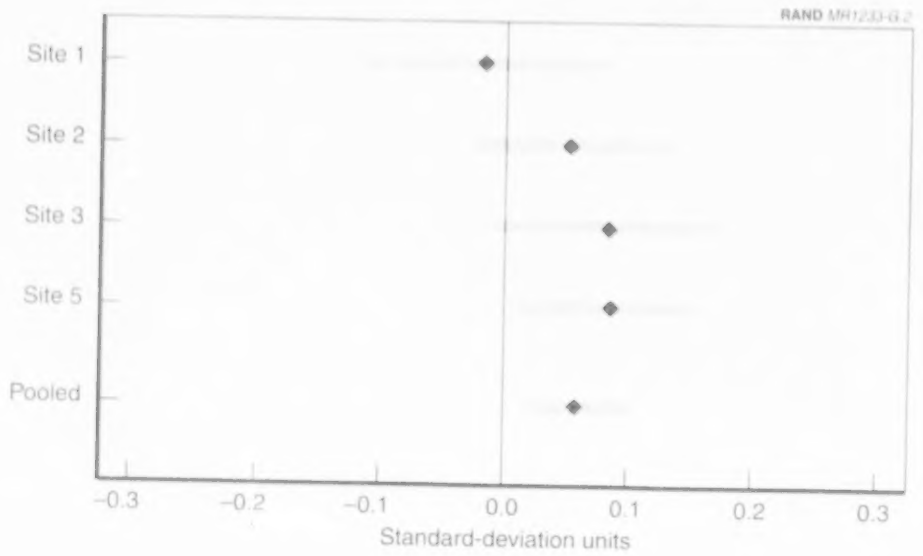


Figure G.2—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Open-Ended Tests

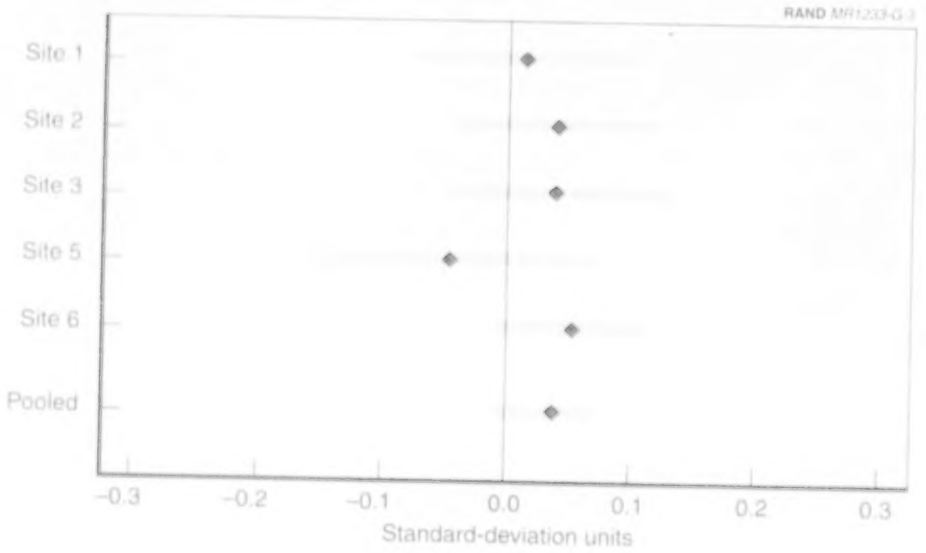


Figure G.3—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Reform Practices, Multiple-Choice Tests

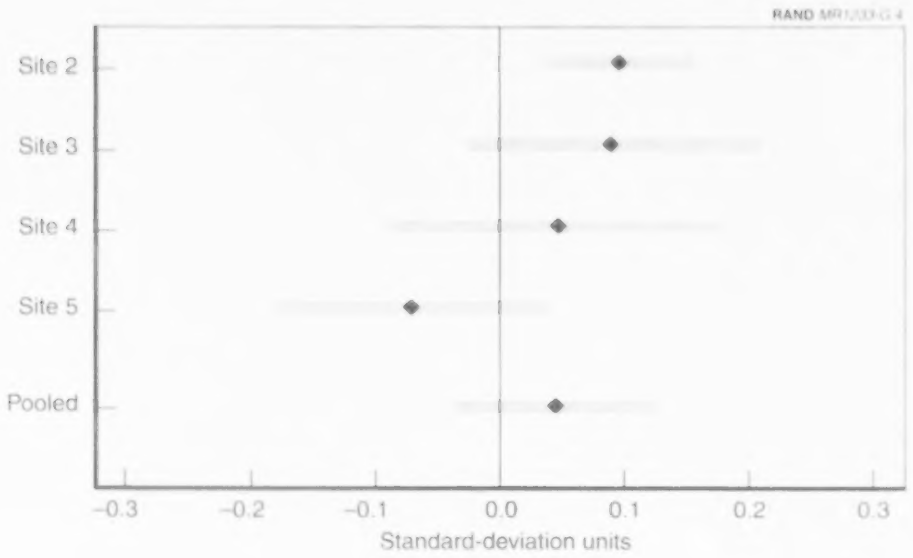


Figure G.4—Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Open-Ended Tests

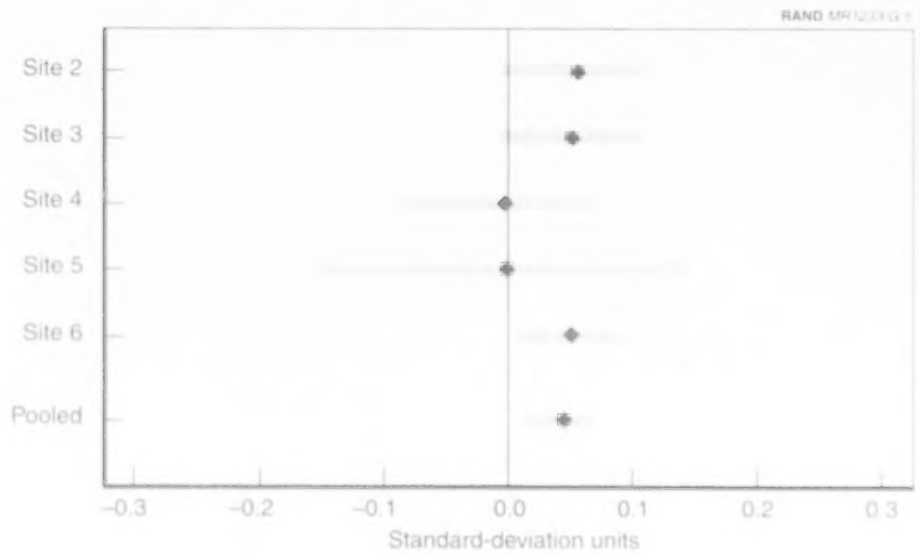


Figure G.5—Estimated Coefficients by Site and Pooled Across Sites for Science: Reform Practices, Multiple-Choice Tests

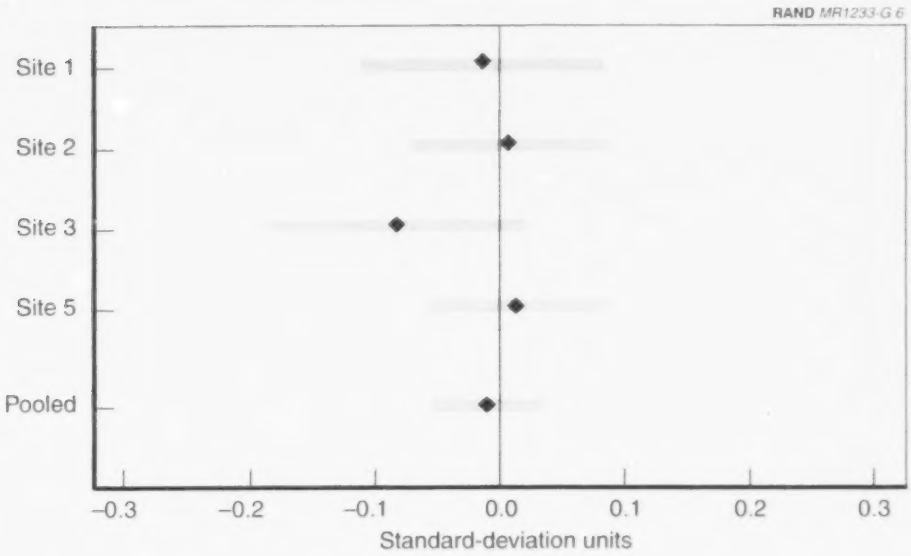


Figure G.6—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Open-Ended Tests

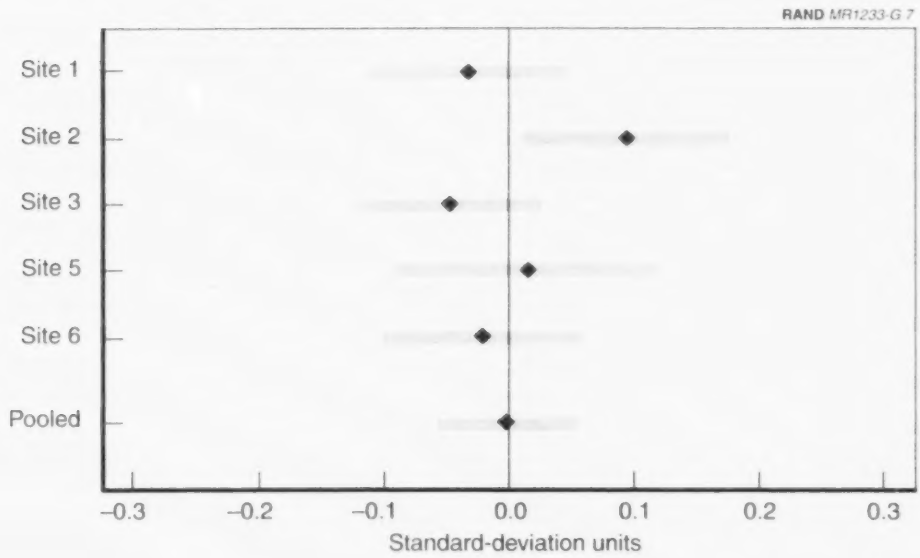


Figure G.7—Estimated Coefficients by Site and Pooled Across Sites for Mathematics: Traditional Practices, Multiple-Choice Tests

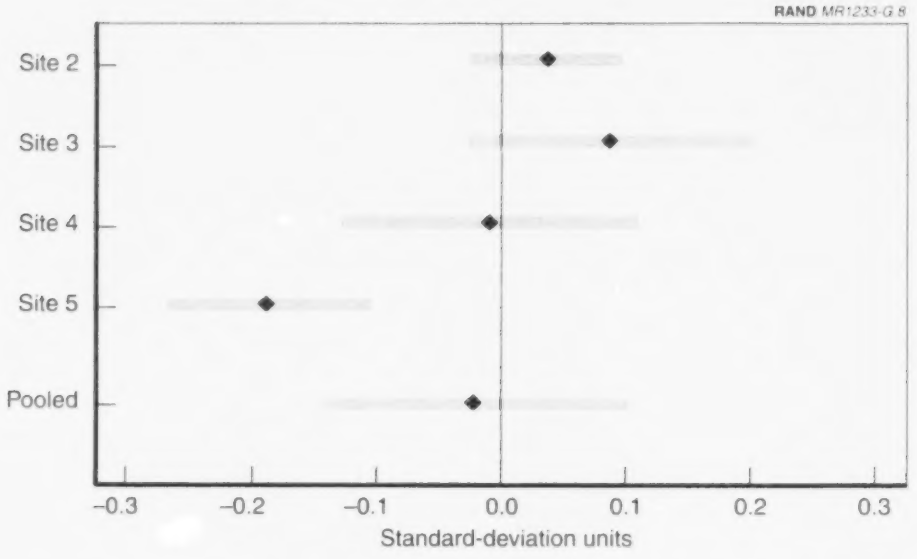


Figure G.8—Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Open-Ended Tests

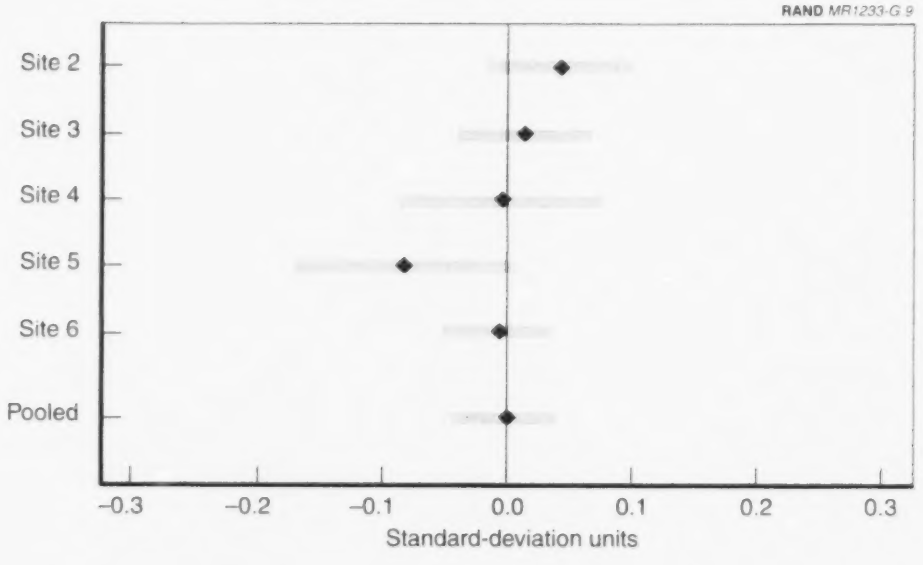


Figure G.9—Estimated Coefficients by Site and Pooled Across Sites for Science: Traditional Practices, Multiple-Choice Tests

Table G.1

Results from Pooled Analyses of Relationships Between Practices and Achievement

Subject	Test Format	Scale	Weighted Average	Standard Error	CI Lower Bound	CI Upper Bound	τ^2	CI for τ^2
Math	OR	Reform	0.059	0.021	0.018	0.099	0.000	(0.000, 0.015)
Math	MC	Reform	0.033	0.018	–0.002	0.068	0.000	(0.000, 0.005)
Science	OR	Reform	0.045	0.040	–0.033	0.123	0.004	(0.000, 0.051)
Science	MC	Reform	0.045	0.013	0.019	0.070	0.000	(0.000, 0.003)
Math	OR	Trad.	–0.009	0.021	–0.049	0.032	0.000	(0.000, 0.012)
Math	MC	Trad.	0.000	0.026	–0.051	0.051	0.002	(0.000, 0.018)
Science	OR	Trad.	–0.020	0.060	–0.138	0.099	0.012	(0.002, 0.123)
Science	MC	Trad.	0.001	0.019	–0.036	0.039	0.001	(0.000, 0.018)

REFERENCES

American Association for the Advancement of Science (1993). *Benchmarks for science literacy: Project 2061.* New York: Oxford University Press.

Bond, L. A., Braskamp, D., and van der Ploeg, A. (1996). *State student assessment programs data base.* Oak Brook, IL: North Central Regional Educational Laboratory.

Cohen, D. K., and Ball, D. L. (1990). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis, 12,* 331–338.

Cohen, D. K., and Hill, H. C. (1998). *State policy and classroom performance: Mathematics reform in California* (CPRE Policy Brief). Philadelphia: Consortium for Policy Research in Education.

Consortium for Policy Research in Education (1995a). *Reforming science, mathematics, and technology education: NSF's State Systemic Initiatives* (CPRE Policy Brief). New Brunswick, NJ: Author.

Consortium for Policy Research in Education (1995b). *Tracking student achievement in science and math: The promise of state assessment programs* (CPRE Policy Brief). New Brunswick, NJ: Author.

Corcoran, T. B., Shields, P. M., and Zucker, A. A. (1998). Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: The SSI's and professional development for teachers. Menlo Park: SRI International.

Fantuzzo, J. W., King, J. A., and Heller, L. R. (1992). Effects of reciprocal peer tutoring on mathematics and school adjustment: A component analysis. *Journal of Educational Psychology, 84,* 331–339.

Fox, J. (1998). NSF programs attacked as weak, unclear. *Education Daily,* July 24, 1–2.

Ginsburg-Block, M. D., and Fantuzzo, J. W. (1998). An evaluation of the relative effectiveness of NCTM standards-based interventions for low-achieving urban elementary students. *Journal of Educational Psychology, 90,* 560–569.

Goldstein, H. (1995). Multilevel statistical models, second edition. London: Arnold.

Greenwood, C. R., Carta, J. J., and Hall, R. V. (1988). The use of peer tutoring strategies in classroom management and educational instruction. *School Psychology Review, 17,* 258–275.

Hill, P. T. (1994). *Reinventing public education* (MR-312-LE/GGF). Santa Monica, CA: RAND.

Knapp, M. S. (1997). Between systemic reforms and the mathematics and science classroom: The dynamics of innovation, implementation, and professional learning. *Review of Educational Research, 67,* 227–266.

Koretz, D., and Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)* (MR-1014-EDU). Santa Monica, CA: RAND.

Koretz, D., Linn, R. L., Dunbar, S. B., and Shepard, L. A. (1991). The effects of high-stakes testing: Preliminary evidence about generalization across tests, in R. L. Linn (chair), *The effects of high stakes testing,* symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Laguarda, K. G. (1998). *Assessing the SSIs' impacts on student achievement: An imperfect science.* Menlo Park, CA: SRI International.

Mayer, D. P. (1998). Do new teaching standards undermine performance on old tests? *Educational Evaluation and Policy Analysis, 20,* 53–73.

McCaffrey, D., and Bell, R. (1997). "Bias reduction in standard error estimates for regression analyses from multi-stage designs with few primary sampling units." Paper presented at the Joint Statistical Meetings, Anaheim CA.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Research Council (1996). *National science education standards.* Washington, DC: National Academy Press.

New Jersey SSI Proposal (1992). *Achieving excellence in mathematics, science and technology education. Statewide systemic initiative: New Jersey proposal to the National Science Foundation.*

Schafer, J. L. (1997). *Imputation of missing covariates under a general linear mixed model.* Technical report available at http://www.stat.psu.edu/~jls/.

Searle, S. R., Cassella, G., and McCullogh, C. E. (1992). *Variance components.* New York: John Wiley and Sons, Inc.

Shields, P. M., Corcoran, T. B., and Zucker, A. A. (1994). *Evaluation of NSF's Statewide Systemic Initiatives (SSI) program: First-year report.* Menlo Park, CA: SRI International.

Shields, P. M., Marsh, J. A., and Adelman, N. E. (1998). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: The SSI's impacts on classroom practice.* Menlo Park: SRI International.

Smith, M., and O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman and B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233–268). Bristol, PA: The Falmer Press.

Stecher, B. M., and Klein, S. P. (Eds.) (1996). *Performance assessments in science: Hands-on tasks and scoring guides* (MR-660-NSF). Santa Monica, CA: RAND.

Tyack, D., and Cuban, L. (1995). *Tinkering toward utopia.* Cambridge, MA: Harvard University Press.

Verschaffel, L., and De Corte, E. (1997). Teaching realistic mathematical modeling in the elementary school: A teaching experiment with fifth-graders. *Journal for Research in Mathematics Education, 28,* 577–601.

Webb, N. M., and Palincsar, A. S. (1996). Group processes in the classroom. In D. C. Berliner and R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 841–873). New York: Macmillan.

Weiss, I. R., Montgomery, D. L., Ridgway, C. J., and Bond, S. L. (1998). *Local Systemic Change through teacher enhancement: Year three cross-site report.* Chapel Hill, NC: Horizon Research, Inc.

Williams, L. (1998). *The Urban Systemic Initiatives (USI) program of the National Science Foundation: Summary update.* Washington, DC: NSF.

Teaching Practices and Student Achievement

In our increasingly technological society, improving students' performance in mathematics and science has become a critical challenge. During the 1990s, the National Science Foundation funded a series of systemic initiatives designed to change the way these subjects are being taught in schools throughout the country. These initiatives sought to align all aspects of the educational system in support of ambitious curriculum and performance standards, with particular emphasis on teacher training and professional development to promote effective changes in instructional practices.

States, urban school districts, and consortia received programs to implement reforms that were consistent with NSF's goals, and in 1996, RAND undertook a study to investigate the relationship between student achievement in mathematics and science and the use of these new instructional practices. We examined six sites that were implementing systemic reforms during the 1996–97 school year and compared them with three comparable non-reform schools. This report presents the results of the grade-level data from the first year of the study.

Our findings provide some, albeit weak, support for the hypothesis that the reform instructional practices are associated with improved student achievement in both mathematics and science. However, as with most large-scale field studies, there are many factors that may have artificially increased or decreased the observed effect sizes. Nevertheless, the consistency of the results across sites is encouraging. Data from the second year of the study will provide additional evidence to aid in the interpretation of these findings.

ISBN 0-8330-2879-0

BEST COPY AVAILABLE

MR-1233-EDU

104